Efficient Robust Estimation of Regression Models (Revision of DP 2006-08)

Cizek, P.

Publication date:
2007

No. 2007–87

EFFICIENT ROBUST ESTIMATION OF REGRESSION MODELS

By Pavel Čížek

October 2007

This is a revised version of CentER Discussion Paper
No. 2006-08

February 2006

ISSN 0924-7815

Efficient robust estimation of regression models*

Pavel Čížek

Department of Econometrics & OR, Faculty of Economics and Business Administration,

Tilburg University, P.O.Box 90153, 5000LE Tilburg, The Netherlands.

Abstract

This paper introduces a new class of robust regression estimators. The proposed two-step least weighted squares (2S-LWS) estimator employs data-adaptive weights determined from the empirical distribution, quantile, or density functions of regression residuals obtained from an initial robust fit. Just like many existing two-step robust methods, the proposed 2S-LWS estimator preserves robust properties of the initial robust estimate. However contrary to existing methods, the first-order asymptotic behavior of 2S-LWS is fully independent of the initial estimate under mild conditions; most importantly, the initial estimator does not need to be $\sqrt{n}$ consistent. Moreover, we prove that 2S-LWS is asymptotically normal under β-mixing conditions and asymptotically efficient if errors are normally distributed. A simulation study documents these theoretical properties in finite samples; in particular, the relative efficiency of 2S-LWS can reach 85–90% in samples of several tens of observations under various distributional models.

Keywords: asymptotic efficiency, breakdown point, least weighted squares

JEL classification: C13, C20, C21, C22

1 Introduction

In statistics and econometrics, more and more attention is paid to techniques that can deal with data containing atypical observations, which can arise from outliers, miscoding, or heterogeneity not captured or presumed in a model. This is of very high importance especially in (non)linear

*This research was supported by the grant GA402/06/0408 of GA ČR.

regression models and time series as the least squares (LS) and maximum likelihood (MLE) estimators are heavily influenced by data contamination. For example using real economic data, Balke and Fomby (1994) document presence of outliers in macroeconomic time series and Sakata and White (1998) or van Dijk et al. (1999) evidence data contamination in financial time series and its adverse effects on estimators (e.g., quasi-maximum likelihood) and tests, respectively. On the other hand, the use of methods robust to atypical observations is infrequent and usually limited to detection of outliers even in recent applications (e.g., Temple, 1998; Woo, 2003), although exceptions exist (e.g., Preminger and Franck, 2007). The reasons could range from missing particular results regarding robust inference, low relative efficiency of many robust methods, or the necessity to choose auxiliary tuning parameters without rigorous guidance. In addition, even the straightforward detection of outliers by a robust method or eye-balling and, after removing outliers, subsequent application of a standard method such as least squares is not a theoretically justified inference method as the usual standard errors (and statistics based on them) will be biased as discussed in the following paragraphs.

To address these issues, we propose a new class of robust estimation methods, the two-step least weighted squares (2S-LWS), which relies on an initial robust estimate and preserves its robust properties. Contrary to existing methods, 2S-LWS has however an asymptotic distribution independent of the initial robust estimation, is asymptotically efficient under normality, and can be free of auxiliary tuning parameters. Most importantly, the asymptotic distribution independent of the initial robust estimate guarantees that correct inference is possible irrespective of the properties of the initial estimator and that the quality of 2S-LWS estimation is not affected by the initial estimator. Consequently, the initial estimator can be chosen to be as robust as possible without concerns about its other qualities and fine-tuning its parameters. To quantify the global robustness of an estimator against large errors and data contamination, one can use, for example, the breakdown point, which measures the smallest contaminated fraction of a sample that can arbitrarily change the estimates (see Section 4 for definition and Rousseeuw, 1997, and Genton and Lucas, 2003, for details). Thus, we concentrate here on the robust methods that achieve the maximum asymptotic breakdown point $1/2$ (in contrast, least squares have its asymptotic breakdown point equal to zero in usual regression settings). Additionally, note that we focus on estimation in the linear regression model, although there are many straighforward extensions to other methods and models as indicated later.

There is a number of high breakdown-point methods, which are insensitive to deviations from the regression model. Most of these methods however pay for their robustness by low relative efficiency in non-contaminated data, especially in normally distributed data. The first equivariant regression estimator with the breakdown point asymptotically equal to $1/2$ was the least median of squares (LMS; Rousseeuw, 1984), which converges only at rate $n^{-1/3}$ (Davies, 1990). Subsequently proposed least trimmed squares (LTS; Rousseeuw, 1985) and S-estimators (Rousseeuw and Yohai, 1984) achieve the usual $\sqrt{n}$ consistency, but they cannot achieve simultaneously high breakdown point and high relative efficiency (Hössjer, 1992). Robust regression methods that can achieve high relative efficiency and maximum breakdown point simultaneously are MM-estimators (Yohai, 1987) and τ-estimators (Yohai and Zamar, 1988). Achieving high relative efficiency of MM- and τ-estimators while preserving the breakdown point $1/2$ is however accompanied by a sizable increase of their bias; moreover, the full efficiency and positive breakdown point cannot be reached at the same time.

To improve the quality of estimation of high breakdown-point methods, Rousseeuw and Leroy (1987) initially suggested to use weighted least squares (WLS), where observations with (robustly-estimated) residuals beyond some fixed cut-off point are assigned zero weight. (This is in spirit similar to one-step M-estimation, see Simpson et al., 1992, and Welsh and Ronchetti, 2002.) Even though this reduces the variability of estimates compared to the initial robust fit, the WLS method cannot improve the convergence rate of the initial robust estimator (He and Portnoy, 1992), and even if the initial estimator is $\sqrt{n}$ consistent, the asymptotic distribution of WLS will depend on the initial robust fit (Welsh and Ronchetti, 2002). Therefore, Gervini and Yohai (2002) proposed to use the WLS strategy with a data-dependent cut-off point. This approach results in a robust and efficient weighted least squares (REWLS) estimator that is asymptotically efficient if errors are normally distributed. Apart from this optimal case of Gaussian data, when REWLS becomes equivalent to the standard least squares (LS) method, the asymptotic distribution of REWLS still depends on the initial estimator (in a known or unknown way depending on the asymptotic behavior of the initial estimator).

In this paper, we propose a new class of high breakdown-point estimation methods, 2S-LWS, which is also based on and improves upon an initial robust estimate. Similarly to Gervini and Yohai (2002), we construct data-adaptive weights using the empirical distribution, or alternatively, using quantile and density functions of the regression residuals obtained from the

initial robust fit. Instead of WLS, we however employ these weights in the context of the least weighted squares (LWS) estimator, which was proposed by Víšek (2002a,b) as a generalization of LTS. To provide an alternative to hard-rejection weights described above for WLS and REWLS, we additionally propose several weighting schemes with strictly positive weights for all observations, which lead to significant improvement in the relative efficiency of the method, especially in small samples.

In comparison to the existing methods, the main benefits of the proposed 2S-LWS method are, apart from its robust properties: (i) the first-order asymptotic independence of the initial estimator, which has to converge only at rate $n^{-\delta}, \delta > 0$, for any underlying distribution; (ii) known asymptotic distribution under mild β-mixing conditions, even for initial estimators that are not $\sqrt{n}$ consistent; and (iii) asymptotic efficiency in the normal model and smaller variance compared to LS for non-Gaussian designs (both asymptotically and in finite samples). In particular, points (i) and (ii) allow us to use as an initial estimator any robust estimator as well as methods based on nonparametric smoothing. Moreover, the principle of 2S-LWS is straightforward to generalize to robust nonlinear regression, instrumental-variables regression (cf. Víšek, 2006), and maximum likelihood estimation (e.g., using Čížek, 2007).

The rest of this paper is organized as follows. The REWLS and 2S-LWS estimators are defined in Sections 2 and 3. Next, the robust and asymptotic properties of 2S-LWS are studied in Sections 4 and 5, respectively. The finite-sample properties of the proposed method, including its relative efficiency, are evaluated and compared with existing methods using Monte Carlo experiments in Section 6. Proofs are given in the appendices.

2 Least weighted squares and efficient robust estimation

Let us consider the linear regression model ($i = 1, \ldots, n$)

$$y_i = x_i^\top \beta^0 + \varepsilon_i, \tag{1}$$

where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ represent the response and explanatory variables and $\beta^0 \in \mathbb{R}^p$ is the underlying value of p unknown regression parameters.

Rousseeuw (1985) proposed to robustly estimate this model by the least trimmed squares

(LTS) estimator,

$$\hat{\beta}_n^{(LTS)} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{h} r_{[i]}^2(\beta), \tag{2}$$

where $r_{[i]}^2(\beta)$ represents the ith order statistics of squared regression residuals $r_1^2(\beta), \ldots, r_n^2(\beta)$ and $r_i(\beta) = y_i - x_i^\top \beta$. The trimming constant h, $\frac{n}{2} < h \leq n$, determines the breakdown point of LTS since definition (2) implies that $n - h$ observations with the largest residuals do not directly affect the estimator. The maximum breakdown point equals asymptotically $1/2$ and is attained for $h = [n/2] + [(p+1)/2]$ (Rousseeuw and Leroy, 1987), whereas for $h = n$, which corresponds to LS, the breakdown point is asymptotically 0.

To improve upon LTS, Víšek (2002a,b) studied a weighted form of LTS, least weighted squares (LWS), which can be defined by

$$\begin{aligned} \hat{\beta}_n^{(LWS)} &= \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} w\left(\frac{2i-1}{2n}\right) r_{[i]}^2(\beta) \\ &= \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} w\left[G_n\{r_i^2(\beta)\} - \frac{1}{2n}\right] r_i^2(\beta), \end{aligned} \tag{3}$$

where $w : \langle 0, 1\rangle \to \mathbb{R}_0^+$ is a weight function and G_n denotes the empirical distribution function of squared residuals $r_i^2(\beta)$. Note that both LS and LTS are special cases of (3) for $w(t) = 1$ and $w(t) = I(t \leq c)$, respectively, for $t \in \langle 0, 1\rangle$ and $c = h/n$. A crucial distinction between LWS and the weighted least squares (WLS) is that weights are assigned to the residual order statistics rather than directly to individual residuals. As the second expression in (3) illustrates, the weight function thus operates on the distribution function of regression residuals.

The LWS estimator in the cross-sectional linear regression was extensively studied by Mašíček (2004) who derives its robust properties, asymptotic normality, and the optimal choice of weight function w provided that the distribution function of the error term ε_i in (1) is known. However, to achieve (asymptotically) breakdown point $1/2$, one has to trim the same amount of observation as LTS and to set $w(t) = 0$ for $t > 0.5$. For Gaussian data, LTS is even the variance-minimizing method among LWS estimators with breakdown point $1/2$. Hence, both LTS and LWS cannot combine a high breakdown point and a good performance in terms of the estimators' variance: for Gaussian data, the relative asymptotic efficiency of LTS with the maximal breakdown point is only 7%.

As a remedy, Gervini and Yohai (2002) proposed the robust and efficient weighted least

squares (REWLS), a method to adaptively determine the observations that needs to be trimmed and to apply LS to the rest of data. Specifically, given initial estimates $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ of regression parameters and residual variance (e.g., from LTS), one can define for each observation weight

$$w_i = I\{|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_0\} \tag{4}$$

for $t_0 > 0$ and $i = 1, \ldots, n$ and then estimate weighted least squares. The cut-off point t_0 is determined by comparing the distribution functions F^+ and F_0^+ of absolute residuals underlying the data and assumed in the model, respectively:

$$d_0 = \sup_{t \geq c} \left\{ [F_0^+(t) - F^+(t)] I[F_0^+(t) - F^+(t) \geq 0] \right\}, \tag{5}$$

$$t_0 = \min\{t : F^+(t) \geq 1 - d_0\}, \tag{6}$$

where $c = 2.5$, for instance. Thus, d_0 measures the largest discrepancy between F_0^+ and F^+ in the tail of the distributions and the cut-off point t_0 is then $1 - d_0$ quantile of the distribution F^+. In practice, F^+ is unknown and has to be replaced by the empirical distribution function F_n^+ of absolute regression residuals, which leads to data-dependent choices d_n and t_n to be used in (4). The described REWLS estimator combines a high breakdown point and asymptotic efficiency under the normal model. In general, the asymptotic distribution of REWLS however depends on the initial robust estimator in a known way if the initial estimator is $\sqrt{n}$ consistent or in an unknown way if the initial estimator converges at a slower rate than $n^{-1/2}$.

3 Two-step least weighted squares

To eliminate the influence of the initial estimator, and additionally, to allow for slowly converging initial estimators (e.g., LMS or kernel density estimators), we now propose to use the data-dependent weights within the LWS estimator instead of simple WLS. Furthermore, to improve the relative efficiency of the estimator asymptotically and in finite samples, we propose several alternative weighting schemes using strictly positive weights and prove that they guarantee high breakdown point and asymptotic efficiency under the normal model.

Let us now assume that $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ are the initial estimates of regression parameters and residual variance. Given the model (1), the corresponding initial regression residuals are $e_i^0 =$

$r_i^2(\hat{\beta}_n^0) = y_i - x^\top \hat{\beta}_n^0, i = 1, \ldots, n$. For the ith order residual statistics $r_{[i]}^2(\hat{\beta}_n^0)$, we can define weight $w_i = \hat{w}_n\{(2i-1)/(2n)\}$, where $\hat{w}_n$ is a weight function that can generally depend on $\hat{\beta}_n^0$, $\hat{\sigma}_n^0$, and e_i^0, but that is assumed to converge to a piecewise continuous function $w : \langle 0, 1\rangle \to \mathbb{R}_0^+$, $\hat{w}_n(t) \to w(t)$ as $n \to \infty$ for all $t \in \langle 0, 1\rangle$. The two-step least weighted squares (2S-LWS) estimator is then defined as

$$\hat{\beta}_n^{(LWS)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \hat{w}_n \left(\frac{2i-1}{2n} \right) r_{[i]}^2(\beta) \tag{7}$$

$$= \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \hat{w}_n \left[G_n\{r_i^2(\beta)\} - \frac{1}{2n} \right] r_i^2(\beta). \tag{8}$$

A specific feature of the proposed estimator is that the weights modify and apply to the values of the empirical distribution function of the squared regression residuals. Therefore, even though the functions $\hat{w}_n$ and w can be arbitrary, one can expect that the 2S-LWS estimator will achieve a high-breakdown only if $w(t)$ approaches zero as t increases towards 1, at least in the case of heavy-tailed distributions. Moreover, since 2S-LWS corresponds to LWS for data-independent weights (e.g., LS correspond to $\hat{w}_n(t) = w(t) = 1$ and LTS to $\hat{w}_n(t) = w(t) = I(t \leq c)$ for $t \in \langle 0, 1\rangle$ and $c \in \langle 1/2, 1\rangle$), it might also seem that there is asymptotically no difference between 2S-LWS using weights $\hat{w}_n \to w$ and LWS using weights w. The crucial distinction however lies in the fact that the 2S-LWS weight function $\hat{w}_n$ can converge to an unknown function w (e.g., depending on the unknown distribution or density functions of ε_i), whereas LWS can be applied only if weight function w is known.

An example of a data-dependent weight function follows from the hard-rejection weights used for REWLS by Gervini and Yohai (2002), who define them by $w_i = I\{|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_n\}$ using $t_n = \min\{t : F_n^+(t) \geq 1 - d_n\}$, see (4). Since $\hat{\sigma}_n^0$ is constant for all observations, we can define the 2S-LWS using REWLS weights (2S-LWS-R) by the weight function

$$\hat{w}_n^R(t) = I(t < 1 - d_n), \tag{9}$$

which converges to $w^R(t) = I(t < 1 - d_0)$. Constant d_0 is defined in (5) using distribution function F_0^+ of $|\varepsilon_i|$ under assumption $\varepsilon_i \sim N(0, 1)$ and d_n is the corresponding sample analogue defined by (5) using empirical distribution function F_n^+ of $|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0|$ and $\hat{\sigma}_n^0 = [1.4826 \cdot \mathrm{MAD}_{i=1,\ldots,n} r_i^2(\hat{\beta}_n^0)]^2$ (for other possible robust estimates of the residual variance, see

Rousseeuw and Croux, 1993, for instance). Whereas REWLS corresponds to WLS with weights $\{w_i\}_{i=1}^n$, the proposed 2S-LWS-R estimator is in fact the LTS estimator (2) with data-dependent trimming $h_n = [(1-d_n)n] = \sum_{i=1}^n w_i$. In other words, given an initial estimate, REWLS eliminates observations with residuals larger than t_n from the sample and apply LS to the remaining data points, whilst 2S-LWS-R just determines the number $h_n = [d_n n]$ of observations to be trimmed and lets LTS to choose them endogenously. This feature permits to update or change the set of trimmed observations in the second step, which can possibly improve quality estimation if the initial fit is poor.

Using the hard-rejection weights, which are either zero or one, might be appropriate if data come from the normal distribution, but it does not use much information from the initial robust estimation. In general, the initial estimate of regression residuals $\{e_i^0\}_{i=1}^n$ allows us to estimate the cumulative distribution function, quantile function, and probability density function of squared residuals. This knowledge can be used to define weight functions that adapt to the data-generating process so that the variance of estimates is as small as possible for any underlying distribution. In the following paragraphs, we propose three such weighting schemes.

First, since the estimation is based on the least squares criterion, which leads to efficient estimates for Gaussian data, we define weights w^Q so that the weighted residuals are normally distributed at the initial estimate $\hat{\beta}_n^0$. Assuming that (standardized) residuals $r_i(\beta), i = 1, \ldots, n$, are from the standard normal distribution $N(0,1)$ at some $\beta \in \mathbb{R}^p$, the squared residuals $r_i^2(\beta)$ follow the χ_1^2 distribution with one degree of freedom. To achieve this for generally distributed squared residuals $r_i^2(\beta)$ at β, we have to transform them to $F_\chi^{-1}\left[G_\beta\left\{r_i^2(\beta)\right\}\right]$, where G_β is the true distribution of $r_i^2(\beta)$ and F_χ and F_χ^{-1} denote the distribution and quantile functions of the χ_1^2 distribution. Hence, we propose to use for 2S-LWS in (8) the weight function ($\hat{\sigma}_n^0$ is an initial estimate of residual variance)

$$\hat{w}_n^Q(t) = \hat{\sigma}_n^0 \frac{F_\chi^{-1}\left(\max\{t, b_n\}\right)}{\left(G_n^0\right)^{-1}\left(\max\{t, b_n\}\right)}, \tag{10}$$

where G_n^0 and $\left(G_n^0\right)^{-1}$ are the empirical distribution and quantile functions of squared residuals $r_i^2\left(\hat{\beta}_n^0\right)$, respectively, that is, $\left(G_n^0\right)^{-1}(t) = r_{[tn]+1}^2(\hat{\beta}_n^0)$ for $t \in (0,1)$, and $b_n = \min\{m/n : r_{[m]}^2(\hat{\beta}_n^0) > 0\}$ is defined and used to avoid dividing by zero. The estimator corresponding to

a robust variance estimate $\hat{\sigma}_n^0 = [1.4826 \cdot \text{MAD}_{i=1,\dots,n} r_i^2(\hat{\beta}_n^0)]^2$ is referred to as 2S-LWS-Q. The weights $\hat{w}_n^Q$ should converge to $w^Q(t) = \sigma^2 F_\chi^{-1}(t)/G_{\beta^0}^{-1}(t)$ under suitable regularity conditions (see Section 5).

The second approach employs the idea that the LS criterion is a finite-sample equivalent of the expectation of squared regression residuals, $\text{E}r_i^2(\beta) = \int r_i^2(\beta)\, g_\beta\{r_i^2(\beta)\} dx d\varepsilon$, where g_β is the density function of $r_i^2(\beta)$. Instead of reweighting residuals to make them normally distributed, we can thus define weights w^D so that, at $\hat{\beta}_n^0$, the expectation of weighted standardized squared residuals $w_i^D r_i^2(\beta)/\sigma^2$ is equal to $1 = \text{E}Z^2$ if $Z \sim N(0,1)$. Denoting the density of the chi-square distribution χ_1^2 by f_χ, the identity

$$\text{E}r_i^2(\beta) = \int r_i^2(\beta) f_\chi\left\{r_i^2(\beta)\right\} dx d\varepsilon = \int r_i^2(\beta) \frac{f_\chi\left\{r_i^2(\beta)\right\}}{g_\beta\{r_i^2(\beta)\}} g_\beta\{r_i^2(\beta)\} dx d\varepsilon$$

indicates that the weight function should be defined by

$$\hat{w}_n^D(t) = \frac{f_\chi\left\{\left(G_n^0\right)^{-1}(t)/\hat{\sigma}_n^0\right\}}{\hat{g}_n^0\left\{\left(G_n^0\right)^{-1}(t)/\hat{\sigma}_n^0\right\}} I\left(\hat{g}_n^0\left\{\left(G_n^0\right)^{-1}(t)/\hat{\sigma}_n^0\right\} > 0\right), \tag{11}$$

where $\hat{g}_n^0$ denotes an estimate of the density function of the initial standardized squared residuals $r_i^2(\hat{\beta}_n^0)/\hat{\sigma}_n^0$ (most density estimates are nonzero at all sample observations). For $\hat{g}_n^0$ being the Rosenblatt-Parzen kernel density estimator with the uniform kernel and bandwidth chosen by Silverman's rule of thumb (Pagan and Ullah, 1999, Chapter 2), the resulting estimator is denoted by 2S-LWS-D. The weights $\hat{w}_n^D$ should converge to $w^D(t) = f_\chi\{G_{\beta^0}^{-1}(t)/\sigma^2\}/g_{\beta^0}\{G_{\beta^0}^{-1}(t)/\sigma^2\}$ on the support of g_{β^0} under regularity conditions (see Section 5).

The third proposal aims at simplifying weights w^D defined in (11), which rely on nonparametric density estimation. For unimodal distributions with relatively light tails, the density functions f_χ and g_β can be approximated by $1 - F_\chi$ and $1 - G_\beta$ in the tails of the respective distributions. Because at least half of observations cannot be trimmed by an equivariant robust estimator, the weight function can be defined, for example, by

$$\hat{w}_n^P(t) = \min\left\{1, \frac{1 - F_\chi\left\{\left(G_n^0\right)^{-1}(t)/\hat{\sigma}_n^0\right\}}{1-t} I(1 > t > 1/2)\right\}. \tag{12}$$

For the initial variance estimate $\hat{\sigma}_n^0 = [1.4826 \cdot \text{MAD}_{i=1,\dots,n} r_i^2(\hat{\beta}_n^0)]^2$, the estimator is referred to

as 2S-LWS-P. The weights $\hat{w}_n^P$ should converge to $w^P(t) = \min\{1, [1 - F_\chi\{G_{\beta^0}^{-1}(t)/\sigma^2\}]/(1-t)\}$ under suitable regularity conditions (see Section 5).

All four proposed weight functions $\hat{w}_n^R, \hat{w}_n^Q, \hat{w}_n^D$, and $\hat{w}_n^P$ allow us to define and compute a 2S-LWS estimate $\hat{\beta}_n^1$ based on an initial estimate $\hat{\beta}_n^0$. First note that the proposals can be further combined and developed; for example, one can use an adaptive cut-off point d_n determined for $\hat{w}_n^R$ to improve $\hat{w}_n^P$ by replacing $I(1 > t > 1/2)$ by $I(1 > t \geq d_n)$ in (12). Next, one can possibly iterate: compute a 2S-LWS estimate based on $\hat{\beta}_n^1$ and so on. The asymptotic results in Section 5 and Monte Carlo experiments in Section 6 however indicate that benefits of such an iterative procedure are negligible both asymptotically and in finite samples.

4 Fundamental properties of 2S-LWS

The four weight functions presented in Section 3 describe four weighted two-step robust estimators based on an initial (high-breakdown point) estimator and LWS. An obvious feature of these estimators is that the weight functions $\hat{w}_n^Q, \hat{w}_n^D$, and $\hat{w}_n^P$ are positive on the whole support of regression residuals and they thus do not reject any observations. The weight functions also do not depend on any auxiliary tuning parameters with the exception of weights $\hat{w}_n^D$ depending on a smoothing parameter and $\hat{w}_n^R$ depending to a small extent on constant c, see (11) and (9), respectively. Nevertheless, the most important properties, which are to be proved here, include asymptotic equivalence of the objective functions of 2S-LWS and LS for Gaussian data and the (high) positive breakdown point of the proposed method. The asymptotic distributions of LWS and 2S-LWS are studied later in Section 5.

One of reasons motivating REWLS and 2S-LWS was the lack of efficiency of many high breakdown-point estimators in models with Gaussian errors. To explain how 2S-LWS improves upon this, we show now that the proposed weight functions functions $\hat{w}_n^R, \hat{w}_n^Q, \hat{w}_n^D$, and $\hat{w}_n^P$ pointwise converge to a constant function on $(0, 1)$ as $n \to \infty$ if $\varepsilon_i \sim N(0, \sigma^2)$ in (1). Hence for normal data, the objective function of 2S-LWS becomes asymptotically identical to the LS criterion. (Note that the following theorem and its proof hold also under more general Assumption A introduced later in Section 5.)

Theorem 4.1 *Assume that $(x_i, \varepsilon_i)_{i=1}^n$ in (1) forms a sequence of independent and identically distributed random variables, $\varepsilon_i \sim N(0, \sigma^2)$, and that the initial estimators $\hat{\beta}_n^0$ of regression*

parameters in (1) and $\hat{\sigma}_n^0$ *of residual variance* $\sigma^2 = \mathrm{var}\varepsilon_i$ *are consistent,* $\hat{\beta}_n^0 \to \beta^0$ *and* $\hat{\sigma}_n^0 \to \sigma^2$ *in probability as* $n \to \infty$*. Then for all* $t \in (0,1)$*, it holds that*

$$\lim_{n\to\infty} \hat{w}_n^R(t) = 1, \lim_{n\to\infty} \hat{w}_n^Q(t) = 1, \lim_{n\to\infty} \hat{w}_n^D(t) = 1, \quad \textit{and} \quad \lim_{n\to\infty} \hat{w}_n^P(t) = 1.$$

Another feature of the proposed 2S-LWS estimator is that it either trims only a (small) adaptively chosen proportion of observations (2S-LWS-R analogously to REWLS), or alternatively, it does not trim observations from its objective function at all, just downweights them (2S-LWS-Q, -D, -P). We have to prove though that this feature does not eliminate or diminish the robust properties of an initial estimator. Intuitively, the use of strictly positive weights does not influence the robust properties to a large extent if the weights decrease sufficiently fast with the degree of outlyingness of an observation. The most slowly decreasing weights here are in most cases weights $\hat{w}_n^Q$ defined in (10), which are indirectly proportional to the value of squared residuals.

To formulate and prove a result concerning the breakdown properties of 2S-LWS, we have to introduce a formal definition of the breakdown point. For the sake of simplicity, we consider independent and identically distributed observations $(y_i, x_i)_{i=1}^n$ (the breakdown point under dependence is generally model-specific; see Genton and Lucas, 2003). The finite-sample breakdown point of an linear-regression estimator $\hat{\beta}_n = T\{(y_i, x_i)_{i=1}^n\}$ can be then defined as (Rousseeuw and Leroy, 1987)

$$\varepsilon_n^*(T) = \frac{1}{n} \max_{m\geq 0} \left\{ m : \max_{I_m=\{i_1,\ldots,i_m\}} \sup_{\tilde{y}_{i_1},\ldots,\tilde{y}_{i_m};\tilde{x}_{i_1},\ldots,\tilde{x}_{i_m}} \left\| T\left\{ (y_i, x_i)_{i\in\{1,\ldots,n\}\setminus I_m} ; (\tilde{y}_i, \tilde{x}_i)_{i\in I_m} \right\} \right\| < \infty \right\}.$$

In other words, it is the maximum number m of observations that can be replaced by arbitrary values $\tilde{y}_i, \tilde{x}_i, i \in I_m$, without making the estimate infinite, that is, completely uninformative and deterministic under contamination. The asymptotic breakdown point of the estimator T is then the corresponding limit, $\varepsilon^*(T) = \lim_{n\to\infty} \varepsilon_n^*(T)$, providing it exists. Further, the breakdown point of a scale estimator $\hat{\sigma}_n = S\{(y_i, x_i)_{i=1}^n\}$ can be defined analogously with the only change that the estimates under contamination must be bounded away both from zero and infinity (in general, the breakdown of an estimator can be generally described as the collapse of the estimator's distribution function to a degenerate one; Genton and Lucas, 2003).

Now, we show that the breakdown point of 2S-LWS with weights $\hat{w}_n^R$, $\hat{w}_n^Q$, $\hat{w}_n^D$, and $\hat{w}_n^P$

equals the minimum of the breakdown points of the initial estimators $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$.

Theorem 4.2 *Let $(y_i, x_i)_{i=1}^n$ be a sequence of independent and identically distributed random vectors, which are almost surely in a general position for $n > p$. Further, let ε_n^{0*} be the finite-sample breakdown point of an initial estimator $(\hat{\beta}_n^0, \hat{\sigma}_n^0)$ of regression parameters and residual variance with limit $\varepsilon^{0*} = \lim_{n\to\infty} \varepsilon_n^{0*}$. Then the finite-sample breakdown points of the 2S-LWS-R, 2S-LWS-Q, 2S-LWS-D, and 2S-LWS-P estimators are larger than or equal to $\min\{\varepsilon_n^{0*}, \{[(n+1)/2] - (p+1)\}/n\}$ and tend to ε^{0*} asymptotically.*

In Theorem 4.2, we limit ourselves only to independent observations so that the intuitive traditional definition of the breakdown point holds. Under dependence, deriving exact breakdown-point results is rather complex and might depend on a specific model. See Sakata and White (2001) and Genton and Lucas (2003), who indicate that the breakdown point ε_n^* of an estimator in cross-sectional regression becomes approximately $\varepsilon_n^*/(1+L)$ in time-series models with at most the Lth lagged variable.

Finally, let us mention that we do not derive here other robust characteristics of 2S-LWS such as influence function (Hampel et al., 1986) because the first-order asymptotic equivalence to LWS proved in the following Section 5 indicates that existing results for LTS (Tableman, 1994) and LWS (Mašíček, 2004) apply.

5 Asymptotics of 2S-LWS

In this section, we first introduce the assumptions necessary for proving the main asymptotic results. Later, the asymptotic distribution of LWS and 2S-LWS is derived and a consistent estimator of the asymptotic covariance matrix is proposed.

5.1 Assumptions

Let us now introduce some notation and definitions. First, the distribution functions of ε_i and ε_i^2 in model (1) are referred to as F and G, respectively, their density functions are denoted f and g, provided that they exist, and the corresponding quantile functions are F^{-1} and G^{-1}, respectively.

Further, let us introduce the concept of β-mixing, which is central to the distributional assumptions made here. A sequence of random variables $\{x_i\}_{i\in\mathbb{N}}$ is said to be absolutely regular

(or β-mixing) if

$$\beta_m = \sup_{i\in\mathbb{N}} \mathrm{E}\left\{ \sup_{B\in\sigma_{i+m}^f} |P(B|\sigma_i^p) - P(B)| \right\} \to 0$$

as $m \to \infty$, where σ-algebras $\sigma_i^p = \sigma(x_i, x_{i-1}, \ldots)$ and $\sigma_i^f = \sigma(x_i, x_{i+1}, \ldots)$; see Davidson (1994) for details. Numbers $\beta_m, m \in \mathbb{N}$, are called mixing coefficients. For example, a stationary ARMA process with continuously distributed innovations is absolutely regular (Mokkadem, 1988).

Now, the assumptions necessary to derive the asymptotic normality of LWS concern the random variables x_i and ε_i in model (1) and weight function w.

Assumption *A*

A1 Random vectors $(x_i, \varepsilon_i)_{i\in\mathbb{N}}$ form a weakly stationary absolutely regular sequence with mixing coefficients β_m satisfying

$$m^{r/(r-2)}(\log m)^{2(r-1)/(r-2)}\beta_m \to 0$$

as $m \to \infty$ for some $r > 2$ and have finite rth moments. Moreover, let $\mathrm{E}(x_i x_i^\top) = Q$ be a nonsingular matrix and

$$n^{-1/4} \max_{i=1,\ldots,n} \|x_i\| = \mathcal{O}_p(1). \tag{13}$$

A2 Let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be a sequence of symmetrically and identically distributed random variables with finite second moments, $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$, and additionally, let ε_i and x_i be independent. The distribution function F of ε_i is absolutely continuous and its probability density function f is assumed to be bounded and continuously differentiable.

A3 Let $w : \langle 0, 1\rangle \to \mathbb{R}_0^+$ be a non-negative, bounded, and left-continuous function that has a bounded derivative everywhere except for a finite set $D = \{d_1, \ldots, d_J\}$ of points of discontinuity. Thus, w can be decomposed to $w = w_s + w_c$, where w_s is a step function and w_c is a continuous and differentiable function.

The first part of Assumption A1 formulates standard conditions of the (uniform) central limit theorem (e.g., Andrews, 1993). For independent and identically distributed (x_i, ε_i), the existence of finite second moments is sufficient, $r = 2$. If only consistency is required, the existence

of first moments of x_i is sufficient (see Čížek, 2006). Further, condition (13) is necessary for the proof of asymptotic normality because of the (possible) discontinuity of the weight function w, and consequently, of the (2S-)LWS objective function (a nonrandom version of this assumption was used for the first time by Jurečková, 1984). Apparently, this condition does not affect random variables with a finite support at all. Čížek (2006, Proposition 1) proves that (13) holds even for random variables with finite second moments and distribution functions having polynomial tails. As the existence of finite second moments is implied by Assumption A1, (13) should not pose a considerable restriction on the explanatory variables. Finally, note that the assumption of random carriers for all variables is made for the sake of simplicity and the results apply in the presence of deterministic variables as well.

Assumption A2 presents standard assumptions on the error term ε_i and its distribution, although they are more restrictive than necessary for the sake of simplicity. For example, the existence of finite second moment can be relaxed if only consistency of (2S-)LWS is required. Furthermore, random variables ε_i and x_i do not have to be independent in general, but similarly to Gervini and Yohai (2002), ε_i conditionally on x_i has to be symmetrically distributed (see Čížek, 2006). Symmetry could be relaxed though by replacing a scale statistics used as an objective function of LTS and LWS by a generalized scale statistics as in Croux et al. (1994) and Stromberg et al. (2000). On the other hand, existence of a differentiable density f is necessary and commonly required when the asymptotic behavior of order statistics is analyzed (Stromberg et al., 2000; Zinde-Walsh, 2002).

Finally, Assumption A3 specifies the most general assumptions on the weight function w under which results presented in Section 5.2 hold. Note though that some proofs are done only for stepwise functions $w = w_s$ and general proofs are just indicated to avoid lengthy technical derivations.

5.2 Asymptotic normality

The main asymptotic results concerning LWS, that is, estimator (3) with a fixed deterministic weight function w, are summarized in the following theorem.

Theorem 5.1 *Let Assumption A hold. Then the least weighted squares estimator* $\hat{\beta}_n^{(LWS)}$ *de-*

fined by a weight function w is $\sqrt{n}$-consistent and asymptotically normal,

$$\sqrt{n}\left(\hat{\beta}_n^{(LWS)} - \beta^0\right) \xrightarrow{\mathcal{L}} N(0, V_w) \tag{14}$$

as $n \to \infty$, where the asymptotic covariance matrix equals

$$V_w = \frac{Q^{-1}\mathrm{var}\left[x_1\varepsilon_1 w\left\{G(\varepsilon_1^2)\right\}\right]Q^{-1}}{\left[\int \varepsilon w\left\{G(\varepsilon^2)\right\} f'(\varepsilon)d\varepsilon\right]^2} \tag{15}$$

provided that the denominator is positive.

Theorem 5.1 specifies the asymptotic distribution of LWS and covers LS and LTS as special cases for $w(t) = 1$ and $w(t) = I(t \leq c)$, respectively, for $t \in \langle 0, 1\rangle$: denoting $a = \sqrt{G^{-1}(c)}$, using integration by parts, and Assumption A2, the denominator of V_w can be expressed as

$$\int_{-a}^{a} \varepsilon f'(\varepsilon)d\varepsilon = [\varepsilon f(\varepsilon)]_{-a}^{a} - \int_{-a}^{a} f(\varepsilon)d\varepsilon = 2af(a) - [F(a) - F(-a)]\,, \tag{16}$$

which equals C_λ in Čížek (2006, Theorem 1) and converges to -1 for $c \to 1$ and $a \to \infty$.

On the other hand, the proposed 2S-LWS estimator uses data-dependent weights, which are by definition random. In all studied cases, the weights are based on estimates of the cumulative distribution function, quantile function, or density function of regression residuals and thus converge to specific nonrandom functions, $\hat{w}_n \to w$ for $n \to \infty$. In the following corollary, we show that the asymptotic distribution of 2S-LWS using a random weight function $\hat{w}_n$ is the same as the one in Theorem 5.1 for LWS using the weight function w.

Corollary 5.2 *Let Assumption A hold for a weight function w and the two-step least weighted squares estimator $\hat{\beta}_n^{(2S\text{-}LWS)}$ be defined by a bounded weighting function $\hat{w}_n$ based on initial estimates $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$. Further, assume that, in probability, $\hat{w}_n(t) \to w(t)$ on $t \in \langle 0, 1\rangle$ and $n^{-\alpha}|\hat{w}_n(t) - w(t)| \to 0$ as $n \to \infty$ uniformly on any compact subset of $(0, 1)$, $\alpha > 0$. Then the two-step least weighted squares estimator $\hat{\beta}_n^{(2S\text{-}LWS)}$ is $\sqrt{n}$-consistent and asymptotically normal,*

$$\sqrt{n}\left(\hat{\beta}_n^{(2S\text{-}LWS)} - \beta^0\right) \xrightarrow{\mathcal{L}} N(0, V_w), \tag{17}$$

as $n \to \infty$, where the asymptotic covariance matrix V_w is defined in Theorem 5.1, equation (15).

This result shows that the 2S-LWS estimator follows asymptotically normal distribution independent of the initial estimate under very general conditions on the initial first-stage estimator: the weights determined from the initial estimate have to converge at rate $n^{-\alpha}$ for some $\alpha > 0$. Let us compare this requirement with known convergence rates in the most typical cases. For weights based on the empirical distribution function G_n^0 of squared regression residuals, Gervini and Yohai (2002) proved for independent and identically distributed data that G_n is uniformly $\sqrt{n}$ consistent even if the initial estimator converges only at $n^{-1/4}$ rate (this conclusion can be directly extended to autoregressive models using recent results of Engler and Nielsen, 2007). Since Corollary 5.2 requires uniform convergence only within compact subset of the w domain, these results also apply if the quantile function of initial squared residuals is used to define $\hat{w}_n$. Finally, for weights based on the estimated density function $\hat{g}_n^0$, Einmahl and Mason (2005), for instance, prove that $\hat{g}_n^0$ is $n^{-1/2}h_n^{-1/2}\log h_n$ consistent uniformly on $\mathbb{R}$ and in h_n, where h_n is the bandwidth used for the kernel density estimation. A simple example of regularity conditions, under which the weighting schemes $\hat{w}_n^R, \hat{w}_n^Q, \hat{w}_n^D$, and $\hat{w}_n^P$ satisfy assumptions of Corollary 5.2, is given in the following lemma.

Lemma 5.3 *Assume that $(x_i, \varepsilon_i)_{i=1}^n$ forms a sequence of independent and identically distributed random variables, the distribution function F of ε_i satisfies Assumption A2 and $z^2 f'(z)$ is bounded. Furthermore, let the initial estimate $(\hat{\beta}_n^0, \hat{\sigma}_n^0)$ be n^α-consistent, $\alpha \geq 2/5$, and the kernel density estimate $\hat{g}_n^0$ be defined by a bandwidth h_n and a kernel function K such that $h_n = O(n^{-1/4})$ as $n \to \infty$ and K is a differentiable probability density function with a bounded support. Then*

$$\begin{aligned}
\sup_{t\in\langle a,b\rangle} \left|\hat{w}_n^R(t) - I(t < 1 - d_0)\right| &= \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big), \\
\sup_{t\in\langle a,b\rangle} \left|\hat{w}_n^Q(t) - \sigma^2 F_\chi^{-1}(t)/G^{-1}(t)\right| &= \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big), \\
\sup_{t\in\langle a,b\rangle} \left|\hat{w}_n^D(t) - f_\chi\left\{\frac{G^{-1}(t)}{\sigma^2}\right\}/g\left\{\frac{G^{-1}(t)}{\sigma^2}\right\}\right| &= \mathcal{O}_p\Big(n^{-\frac{1}{2}}h_n^{-1}\Big), \\
\sup_{t\in\langle a,b\rangle} \left|\hat{w}_n^P(t) - \min\left\{1, \frac{1 - F_\chi[G^{-1}(t)/\sigma^2]}{1-t}\right\}\right| &= \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)
\end{aligned}$$

for any $0 < a < b < 1$ as $n \to \infty$, where $\sigma^2 = \mathrm{var}\varepsilon_i$.

Next, to highlight the difference between the existing REWLS and proposed 2S-LWS estimators, let us now compare the asymptotic variances of the two estimators for the hard-rejection weights and independent and identically distributed data. Specifically, we compare REWLS with weights $w^1(t) = I(t < t_0)$ as used in Gervini and Yohai (2002), see (4), and 2S-LWS-R with equivalently defined weights $w^2(t) = I\{G(t^2) < d_0\}$, see (9). Assuming that the initial estimator $\hat{\beta}_n^0$ is $\sqrt{n}$ consistent and admits asymptotic linear expansion

$$\hat{\beta}_n^0 - \beta^0 = \frac{\Gamma_n}{n} \sum_{i=1}^{n} \psi\left(\frac{\varepsilon_i}{\sigma}\right) x_i + o_p\left(n^{-\frac{1}{2}}\right)$$

for a positive definite matrix $\Gamma_n \to \Gamma$ as $n \to \infty$, the asymptotic distribution of REWLS is known and its variance matrix equals

$$V^{(REWLS)} = \frac{Q^{-1}}{\pi_1^2} \mathrm{var}\left\{w^1(|\varepsilon_i|)\varepsilon_i + \frac{\tau_1}{\sigma}\psi(\varepsilon_i)\Gamma\right\} \tag{18}$$

(Gervini and Yohai, 2002, Theorem 4.1), where

$$\pi_k = \int w^k(|\varepsilon|) f(\varepsilon) d\varepsilon, \qquad \tau_k = \pi_k + \int \varepsilon w^k(|\varepsilon|) f'(\varepsilon) d\varepsilon.$$

On the other hand, Theorem 5.1 indicates that the asymptotic variance matrix of 2S-LWS-R equals

$$V^{(2S\text{-}LWS\text{-}R)} = \frac{Q^{-1}}{(\pi_2 - \tau_2)^2} \mathrm{var}\left\{w^2(\varepsilon_i^2)\varepsilon_i\right\}. \tag{19}$$

Using (16), we can express

$$\pi_k = F(t_0) - F(-t_0), \qquad \tau_k = 2t_0 f(t_0).$$

Hence for Gaussian data, weighting functions are constant ($t_0 = \infty$ by Theorem 4.1), $\pi_k = 1$ and $\tau_k = 0$, and variances of both methods are equal, $V^{(REWLS)} = V^{(2S\text{-}LWS\text{-}R)}$.

However, if trimming takes place (e.g., for a heavy-tailed F and G), $\pi_k < 1, \tau_k > 0$, and the asymptotic variance of REWLS depends on the the initial estimator by means of terms $\psi(\varepsilon_i)$ and Γ in (18). In comparison, the variance (19) of 2S-LWS-R stays independent of the initial estimator (this holds even if the initial estimator is not $\sqrt{n}$ consistent). Nevertheless, the asymptotic variance of both estimators is growing as the amount of trimming increases and t_0

decreases because π_k is increasing, τ_k decreasing, and $\pi_k - \tau_k$ increasing in t_0. Obviously as t_0 decreases, the denominator $(\pi_2 - \tau_2)^{-2}$ in variance (19) of 2S-LWS-R increases faster than the denominator π_2^{-2} in variance (18) of REWLS, but this advantage of REWLS is eliminated by the already mentioned second term in variance (18) coming from the initial estimator. Exact comparison thus depends on the initial estimator and underlying data distribution; asymptotic and finite-sample results for various error distributions are presented in Section 6.

Finally, to make the results of Theorem 5.1 and Corollary 5.2 practically applicable, we propose now a consistent and computationally feasible estimator of V_w.

Theorem 5.4 *Let Assumption A hold and decomposition $w = w_s + w_c$ be such that $w_s(1) = 0$. Further, let e_{in} denote the regression residual $r_i(\hat{\beta}_n)$ at the LWS or 2S-LWS estimate, $C_V = \mathrm{var}\left[x_1\varepsilon_1 w\left\{G(\varepsilon_1^2)\right\}\right]$, and $C_I = -\int \varepsilon w\left\{G(\varepsilon^2)\right\} f'(\varepsilon)d\varepsilon \neq 0$. Finally, let $\hat{q}_{jn}^2 = e_{[d_j n]}^2$ be the d_jth empirical quantile of $\{e_{in}^2\}_{i=1}^n$ for $j = 1, \ldots, J$ and denote $d_{J+1} = 1$. The estimator $\hat{V}_{wn} = \hat{Q}_n^{-1}\hat{C}_{Vn}\hat{Q}_n^{-1}/\hat{C}_{In}^2$ is weakly consistent for the asymptotic covariance matrix V_w of $\hat{\beta}_n$, $\hat{V}_{wn} \to V_w$ in probability as $n \to \infty$, where*

- $\hat{Q}_n = \sum_{i=1}^n x_i x_i^\top / n$,
- $\hat{C}_{Vn} = \sum_{i=1}^n x_i^2 e_{in}^2 w^2\left[\hat{G}_n(e_{in}^2) - \frac{1}{2n}\right]/n$,
- $\hat{C}_{In} = \hat{C}_{In}^s + \hat{C}_{In}^c$ *with*

$$\hat{C}_{In}^s = \sum_{j=1}^{J}\{w_s(d_j) - w_s(d_{j+1})\}\left\{d_j - 2\hat{q}_{jn}^2 g(\hat{q}_{jn}^2)\right\} \tag{20}$$

and

$$\hat{C}_{In}^c = \frac{1}{n}\sum_{i=1}^n w_c\left\{\hat{G}_n(e_{in}^2) - \frac{1}{2n}\right\} + \frac{2}{n}\sum_{i=1}^n e_{in}^2 w_c'\left\{\hat{G}_n(e_{in}^2) - \frac{1}{2n}\right\}\hat{g}_n(e_{in}^2), \tag{21}$$

- *$\hat{G}_n$ denotes a uniformly consistent estimator of the distribution function G,*
- *and $\hat{g}_n$ is a uniformly consistent estimator of the density function g.*

Theorem 5.4 does not specify what estimates $\hat{G}_n$ and $\hat{g}_n$ of the distribution and density functions of squared regression residuals should be used. There is however a wide range of estimation methods available: for example, $\hat{G}_n$ can be a standard or smoothed empirical distribution function (Fernholz, 1997); similarly, $\hat{g}_n$ can be represented by a kernel density estimator (see

Pagan and Ullah, 1999, Chapter 2, for an overview). Further, let us note that C_I and both terms (20) and (21) are positive in usual situations. For example for unimodal symmetric distributions, $f'(\varepsilon)$ is an asymmetric function, $\varepsilon f'(\varepsilon) < 0$ is a symmetric function, and $w(t) \geq 0$ implies $C_I > 0$ (see Assumptions A2 and A3). Similarly, Čížek (2006) proved that (16) and its empirical counterpart (20) are positive for unimodal distributions. Term (21) is positive for non-increasing weight functions w, for instance.

6 Finite-sample properties

In this section, we present a Monte Carlo study done to assess finite-sample behavior of the proposed 2S-LWS estimators and to compare it with existing methods. The influence of some initial estimators on REWLS and 2S-LWS are discussed as well. First, the relative efficiency of all methods is examined at various sample sizes and error distributions for cross-sectional data (Section 6.1). Later, we study all estimators under heteroscedasticity, errors from finite-mixtures, and data contamination by outliers as well (Section 6.2). In these simulation experiments, we compare all four proposed variants of 2S-LWS with relevant existing estimators: standard MLE and LS; robust LMS, LTS, and S estimators set up for the maximum possible breakdown point (i.e., LTS with the trimming constant $h = [n/2] + [(p+1)/2]$ and the S estimator with Tukey's biweight function and $c = 1.547$; see Rousseeuw and Leroy, 1987, for details); RDL_1, a robustly weighted L_1 estimator by Hubert and Rousseeuw (1997), designed for models with binary covariates, which also does not fully downweight any observation; and data-adaptive robust REWLS with hard-rejection weights (4) using $c = 2.5$ in (5), see Gervini and Yohai (2002) for details. Unless stated otherwise, all adaptive estimators use for the initial robust fit the described S estimator.

6.1 Finite-sample efficiency

The relative efficiency of an estimator T can be defined as the ratio of the mean squared errors (MSE) of the asymptotically efficient MLE and the respective estimator T. Having an experiment consisting of S simulated samples of size n, we thus have to obtain S MLE estimates $\hat{\beta}_n^{(MLE,s)}$ and S estimates $\hat{\beta}_n^{(T,s)}$, $s = 1, \ldots, S$. The relative mean squared efficiency is

Table 1: Relative MSE efficiencies for normal errors, $\varepsilon_i \sim N(0,1)$.

Estimation	Sample size n					
method	25	50	100	200	400	∞
LS	1.00	1.00	1.00	1.00	1.00	1.00
LMS	0.21	0.18	0.16	0.14	0.11	0.00
LTS	0.22	0.19	0.15	0.12	0.10	0.07
S	0.35	0.29	0.24	0.23	0.22	0.28
RDL_1	0.42	0.50	0.50	0.57	0.57	—
REWLS	0.46	0.63	0.76	0.90	0.91	1.00
2S-LWS-R	0.49	0.65	0.80	0.92	0.92	1.00
2S-LWS-Q	0.74	0.84	0.92	0.97	0.98	1.00
2S-LWS-D	0.45	0.56	0.67	0.85	0.84	1.00
2S-LWS-P	0.44	0.57	0.73	0.88	0.89	1.00

then defined by

$$Eff = \frac{\sum_{s=1}^{S} \left\| \hat{\beta}_n^{(MLE,s)} - \beta^0 \right\|^2}{\sum_{s=1}^{S} \left\| \hat{\beta}_n^{(T,s)} - \beta^0 \right\|^2}.$$

Note that, in this section, the simulated results are complemented by the asymptotic relative efficiencies if they are known and independent of other factors (e.g., the variance of initial estimate).

We evaluate the relative efficiency for the regression model

$$y_i = 0.5 + x_{1i} - 2x_{2i} + \varepsilon_i, \tag{22}$$

where $x_{1i}, x_{2i} \sim N(0,1)$ and $(x_{1i}, x_{2i}, \varepsilon_i)_{i=1}^n$ forms a sequence of independent random vectors. The considered error distributions are the standard normal $\varepsilon_i \sim N(0,1)$, double exponential $\varepsilon_i \sim DExp(1)$, and Student $\varepsilon_i \sim t(5)$ distributions, which are further referred to as NORM, DEXP, and STD(5), respectively, and cover distribution functions with both exponential and polynomial tails. Results for sample sizes from $n = 25$ to 400 are based on 500 simulated samples.

First, let us compare the behavior of all methods for the linear regression model NORM using results in Table 1. The initial robust estimators, LMS, LTS, and S, exhibit common behavior characterized by a low relative efficiency, which is decreasing with the sample size. Only the relative efficiency of RDL_1, which does not fully downweight any observation, increases slightly as the sample size grows. On the other hand, the relative efficiency of the adaptive robust estimators, REWLS and 2S-LWS, significantly increases with the sample size and converges to

Table 2: Relative MSE efficiencies for normal errors, $\varepsilon_i \sim N(0,1)$, as a function of an initial estimator.

Initial estimator	Adaptive estimator	Sample size n					
		25	50	100	200	400	∞
LMS	REWLS	0.28	0.47	0.68	0.74	0.83	1.00
	2S-LWS-Q	0.71	0.82	0.94	0.92	0.97	1.00
LTS	REWLS	0.37	0.56	0.74	0.79	0.85	1.00
	2S-LWS-Q	0.69	0.81	0.93	0.95	0.99	1.00
S	REWLS	0.43	0.62	0.77	0.88	0.90	1.00
	2S-LWS-Q	0.70	0.82	0.94	0.96	0.99	1.00

1 (asymptotically by Theorem 4.1). The relative efficiency of 2S-LWS-R, which uses the same hard-rejection weights as REWLS, is a bit better than REWLS at all sample sizes: the relative efficiencies grow from $(0.63, 0.65)$ at $n = 50$ to $(0.90, 0.92)$ at $n = 400$. The performance of 2S-LWS-D and 2S-LWS-P, which heavily downweight observations with large residuals similarly to 2S-LWS-R, but rely on estimated weights, is slightly worse than that of REWLS and 2S-LWS-R. On the contrary, 2S-LWS-Q achieves a high relative efficiency 0.84 already at rather small sample sizes $n = 50$ and performs almost as well as LS (MLE) at $n = 200$ and $n = 400$. Thus, 2S-LWS-Q is superior to all other robust methods in model NORM.

An additional note concerns the choice of an initial estimation method for REWLS and 2S-LWS. Both methods rely on an initial robust estimator such as LMS, LTS, and S. For Gaussian data, the initial estimate does not influence the variance of REWLS estimates asymptotically, but the (asymptotic) distribution of REWLS depends on the initial estimate if trimming takes place (Gervini and Yohai, 2002). On the other hand, the (asymptotic) distribution 2S-LWS does not depend on the initial estimate irrespective of the underlying distribution and the amount of trimming. In finite samples, the sensitivity of both methods to the choice of an initial estimator is documented in Table 2 for REWLS and 2S-LWS-Q. Obviously, the dependence on the initial estimator is practically negligible in the case of the proposed 2S-LWS method at all sample sizes. On the other hand, REWLS results significantly differ for various initial estimators, although the differences tend to get smaller as the sample size increases; this is consistent with the simulation results of Gervini and Yohai (2002).

Before discussing other simulation experiments, we indicate here by an example how quickly the standard deviations of the 2S-LWS-Q regression estimates, obtained in simulations from (22), converge to their asymptotic values given in Theorem 5.1. Table 3 summarizes the simu-

Table 3: Finite-sample and asymptotic variances of 2S-LWS-Q estimates.

Standard deviations	Sample size				
Parameter	25	50	100	200	400
Intercept	0.252	0.152	0.103	0.080	0.053
Slope x_1	0.253	0.168	0.110	0.070	0.053
Slope x_2	0.263	0.157	0.112	0.074	0.050
Asymptotic	0.200	0.141	0.100	0.070	0.050

Table 4: Relative MSE efficiencies for Student errors, $\varepsilon_i \sim t(5)$.

Estimation	Sample size n					
method	25	50	100	200	400	∞
LS	0.85	0.83	0.80	0.80	0.78	0.80
LMS	0.25	0.24	0.23	0.20	0.17	0.00
LTS	0.27	0.24	0.22	0.19	0.16	0.14
S	0.42	0.39	0.38	0.33	0.30	0.40
RDL_1	0.52	0.57	0.62	0.65	0.63	—
REWLS	0.55	0.71	0.86	0.89	0.91	—
2S-LWS-R	0.57	0.76	0.86	0.88	0.88	0.90
2S-LWS-Q	0.80	0.90	0.96	0.98	0.98	0.98
2S-LWS-D	0.53	0.66	0.82	0.83	0.87	0.97
2S-LWS-P	0.54	0.69	0.83	0.88	0.95	0.97

lated and asymptotic results; note that the asymptotic variances of all regression parameters are equal due to the design of model NORM. Even though the asymptotic variance underestimates the true variance as expected, it seems that the error of the asymptotic approximation does not exceed 10–12% from sample sizes $n = 100$ on and (15) can be thus reasonably used for such samples.

Next, Table 4 summarizes the simulation results for model STD(5). The performance of all initial robust estimators, LMS, LTS, S, and RDL_1, mirrors the behavior in model NORM apart from the fact that the relative efficiencies are slightly higher at all sample sizes. Moreover, LS is now not efficient anymore (its relative efficiency decreases from 0.85 to 0.80), but its performance is significantly better than that of LMS. On the contrary, the relative efficiency of all adaptive robust estimators grows with an increasing sample size. Similarly to model NORM, relative efficiency for most adaptive methods is relatively low at small samples, for example around 0.7 at $n = 50$, but reaches levels around or above 0.90 at larger samples. The method 2S-LWS-Q proves again to be superior to the other ones since it reaches relative efficiency 0.90 already at $n = 50$ and further 0.98 at $n = 400$, being very close to the performance of MLE and outperforming LS at all samples of 50 or more observations.

Table 5: Relative MSE efficiencies for double exponential errors, $\varepsilon_i \sim DExp(1)$.

Estimation	Sample size n					
method	25	50	100	200	400	∞
LS	0.88	0.73	0.65	0.63	0.56	0.50
LMS	0.35	0.36	0.33	0.27	0.22	0.00
LTS	0.40	0.40	0.40	0.38	0.38	0.35
S	0.58	0.58	0.59	0.55	0.41	0.60
RDL_1	0.65	0.77	0.83	0.83	0.82	—
REWLS	0.70	0.81	0.87	0.85	0.73	—
2S-LWS-R	0.69	0.79	0.78	0.77	0.67	0.55
2S-LWS-Q	0.99	1.03	1.01	1.00	0.90	0.76
2S-LWS-D	0.70	0.82	0.84	0.85	0.75	0.63
2S-LWS-P	0.69	0.85	0.90	0.92	0.82	0.70

Finally, we discuss simulation results from model DEXP that are found in Table 5. The most initial (robust) estimators perform similarly as in models NORM and STD(5). The main differences are that the relatively efficiency of RDL_1 now reaches efficiency levels above 0.8 (which is not surprising given that it is based on the weighted least absolute deviation estimator) and that the relative efficiency of LS drops down to 0.56 at larger samples. Contrary to previous simulations, the adaptive methods seem to exhibit relative efficiency above 0.7, which more or less constant or slowly decreasing as n changes from 100 to 400 observations. Additionally, 2S-LWS-P (and partially also 2S-LWS-D) is now preferable to REWLS, which in turn outperforms 2S-LWS-R. The 2S-LWS-Q performs again better than all other methods, with relative efficiencies being above 0.90 all the time and reaching 1.00 at samples with 50 to 200 observations.

Altogether, the only method that achieves in all models relative efficiency above 0.90, at least at larger samples, was the proposed 2S-LWS-Q estimator.

6.2 Behavior for contaminated data

To learn more about finite-sample behavior of the discussed estimators in the presence of heteroscedasticity, outliers, and so on, we again employ model (22) under various distributional schemes and compare different estimators by means of their mean squared errors: $MSE = \frac{1}{S}\sum_{s=1}^{S}\left\|\hat{\beta}_n^{(T,s)} - \beta^0\right\|$, where $\hat{\beta}_n^{(T,s)}, s = 1, \ldots, S$, are the estimates for S simulated samples. We use following data generating processes, where $x_{1i}, x_{2i} \sim N(0,1)$ unless stated otherwise:

NORM: Clean Gaussian data for the reference purpose, $\varepsilon_i \sim N(0,1)$.

Table 6: Mean squared errors of all methods in various cross-sectional regression models, $n = 100$.

Model	Estimation method						
	LS	S	REWLS	2S-LWS			
				R	Q	D	P
NORM	0.032	0.114	0.041	0.040	0.034	0.045	0.042
STD(3)	0.098	0.118	0.054	0.056	0.050	0.056	0.056
MIX	0.096	0.770	0.132	0.134	0.089	0.164	0.214
HET	0.441	0.122	0.090	0.100	0.064	0.90	0.92
OUT(0.10)	2.554	0.112	0.040	0.040	0.042	0.045	0.047
LOUT(0.10,4)	2.756	0.108	0.046	0.046	0.052	0.052	0.053
OUT(0.25)	6.840	0.092	0.045	0.044	0.050	0.055	0.049
LOUT(0.25,6)	3.543	0.101	0.062	0.063	0.088	0.091	0.069
OUT(0.40)	9.986	0.102	0.064	0.064	0.069	0.103	0.062
LOUT(0.40,8)	4.883	0.152	0.119	0.120	0.212	0.190	0.120

STD(d): Data with errors from a heavy-tailed distribution, $\varepsilon_i \sim t(d)$, where $t(d)$ denotes the Student distribution with d degrees of freedom.

MIX: Clean data, where the error term comes from a symmetric mixture of normal distributions, $\varepsilon_i \sim 0.6N(0,1) + 0.2N(-2.5, 0.25) + 0.2N(2.5, 0.25)$.

HET: Data with heteroscedasticity of a known form, $\varepsilon_i \sim N(0, e^{2x_1})$.

OUT(a): Data contaminated by $[an]$ (vertical) outliers, $\varepsilon_i \sim (1-a)N(0,1) + aU(-50, 50)$.

LOUT(a,l): Data contaminated by outliers in a leverage position, where a fraction a of observations satisfies $x_{1i}, x_{2i} \sim N(0,1)$ and $\varepsilon_i \sim N(0,1)$ and the complementary fraction $1-a$ of observations follows $x_{1i}, x_{2i} \sim N(l,1)$ and $\varepsilon_i \sim U(-50, 50)$.

All simulations in this section are done for sample size $n = 100$ and 500 simulated samples. The adaptive estimators are based on an initial S estimate in all cases.

Let us now discuss simulation results summarized in Table 6. The results for models NORM and STD(3) resemble those in Section 6.1: the adaptive methods, REWLS and 2S-LWS, perform well both for light-tailed and heavy-tailed data, with 2S-LWS-Q being the best one. In comparison, LS is preferable in model NORM, but worse than REWLS and 2S-LWS in model STD(3).

To examine the performance of all methods in less standard situations, where the error distribution is not unimodal, model MIX is included. In general, the initial S estimator performs poorly here (the same holds for unreported LMS and LTS). Contrary to the previous simulations,

REWLS and 2S-LWS-R perform now significantly better than 2S-LWS-D and 2S-LWS-P, where the last one is the worst one. This is caused by the fact that the density function $g(z)$ of the squared errors is not monotonically decreasing for $z \to \infty$ as the weighting scheme of 2S-LWS-P assumes. The best methods is again 2S-LWS-Q, which outperforms both robust methods and LS.

A similar experiment involves estimation under heteroscedasticity, see model HET, without actually knowing and modeling heteroscedasticity. In this case, LS becomes inferior to all (adaptive) robust methods. As in previous simulations, REWLS and 2S-LWS-D/P estimate equally well and 2S-LWS-Q outperforms by far all other methods.

Finally, simulations with data contaminated by outliers are carried out, see models OUT(0.10) to LOUT(0.40,8). Although an increasing amount of contamination has an adverse effect on all estimation methods, LS is affected to such an extent that the estimates are useless. All robust estimators keep their MSE relatively small even under extreme levels of contamination. In all cases, adaptive robust methods provide best and most stable estimates. There are however differences among weighting schemes. The best performing methods are REWLS and 2S-LWS-R that assign weight zero to outlying observations. The two methods are closely matched by 2S-LWS-P, which uses weights decreasing very fast with the absolute value of regression residuals. On the other hand, 2S-LWS-D and especially 2S-LWS-Q, which downweights outlying observations as little as possible, exhibit a larger bias and MSE, in particular in models LOUT$(\cdot, \cdot)$ containing leverage points. Nevertheless, 2S-LWS-Q is still preferable to the S estimators except for model LOUT(0.40,8).

To conclude, the 2S-LWS-Q estimator is preferable or at least comparable both to LS and other robust methods under most distributional models. Due to its no-rejection feature, it is more sensitive to data contamination than REWLS, for instance, but the difference is not pronounced unless the contamination level is very high. If this is a concern, REWLS or 2S-LWS-R can be employed.

7 Conclusion

In this paper, a new class of robust estimation methods is introduced, which offers not only robustness in terms of a high breakdown point, but also asymptotic efficiency for Gaussian data and high relative efficiency under many other distributional models both asymptotically and

in finite samples. This is especially the case of the 2S-LWS-Q method. Its only weak point is higher sensitivity (bias) in data with a large proportion of outlying leverage points, in which case REWLS or the proposed 2S-LWS-R estimator form a reasonable alternative.

Although the methods are proposed and discussed in the context of (homoscedastic) linear regression, many extensions are straightforward. This does not only include regression under heteroscedasticity, but also instrumental variable estimation proposed for LWS by Víšek (2006), nonlinear regression using results of Čížek (2006), or maximum likelihood estimation (Čížek, 2007) as long as the response variable is continuous.

A Proofs of the fundamental properties

Proof of Theorem 4.1: Let us consider an arbitrary $t \in (0,1)$ and $z \in \mathbb{R}$. We will first establish consistency of various statistics in (10)–(12) for $n \to \infty$; the result for the weight function $\hat{w}_n^R$ defined in (9) is derived in Gervini and Yohai (2002, Lemma 4.1).

By assumptions of the theorem, $\hat{\beta}_n^0 \to \beta^0$ and $\hat{\sigma}_n^0 \to \sigma$. Hence, residuals $r_i(\hat{\beta}_n^0) \to r_i(\beta^0) \equiv \varepsilon_i$ in probability as $n \to \infty$. Further, $r^2_{[tn]}(\beta) = G_n^{-1}(t) \to G_\beta^{-1}(t)$ uniformly in probability over some neighborhood $U(\beta^0, \delta), \delta > 0$, (Čížek, 2004, Lemma A.2), where G_β and G_β^{-1} are the distribution and quantile functions of $r_i^2(\beta)$. Finally, the Rozenblatt-Parzen estimator $\hat{g}_n^0$ of g_β in (11) is uniformly consistent on $\mathbb{R}$, $\sup_{z\in\mathbb{R}} \left|\hat{g}_n^0(z) - g_\beta(z)\right| \to 0$ in probability as $n \to \infty$ (Castellana and Leadbetter, 1986).

For the weight function $\hat{w}_n^Q(t)$, the assumption $\varepsilon_i \sim N(0,\sigma)$ then implies that $G_n^{-1}(t)/\hat{\sigma}_n^0 = r^2_{[tn]}(\beta)/\hat{\sigma}_n^0 \to F_\chi^{-1}(t)$ in probability. Consequently, $\hat{w}_n^Q(t) \to 1$ in probability as $n \to \infty$.

For the weight function $\hat{w}_n^D(t)$, the uniform consistency of $\hat{g}_n$ and the result $G_n^{-1}(t)/\hat{\sigma}_n^0 \to F_\chi^{-1}(t)$ imply that $\hat{g}_n\left\{G_n^{-1}(t)/\hat{\sigma}_n^0\right\} \to f_\chi\left\{F_\chi^{-1}(t)\right\} > 0$ in probability. Hence, $\hat{w}_n^D(t) \to 1$ in probability as $n \to \infty$.

Finally, using once again $G_n^{-1}(t)/\hat{\sigma}_n^0 \to F_\chi^{-1}(t)$, the continuity of F_χ results in $F_\chi\left\{G_n^{-1}(t)/\hat{\sigma}_n^0\right\} \to t$ and $\hat{w}_n^P(t) \to 1$ in probability as $n \to \infty$. □

Proof of Theorem 4.2: For a given sample $\{y_i, x_i\}_{i=1}^n$ of size n, let $\varepsilon_n^* = \min\{\varepsilon_n^{0*}, \{[(n+1)/2] - (p+1)\}/n\}$. Further, assume that the breakdown point of any proposed 2S-LWS is smaller than ε_n^*, that is, there exist $m \leq n\varepsilon_n^*$, an index set I_m of size m, and sequences of points $\{\tilde{y}_i^s, \tilde{x}_i^s\}_{s\in\mathbb{N}}, i \in I_m$, such that 2S-LWS estimators $\hat{\beta}_n^s$ applied to samples $C_m^s =$

$\{y_i, x_i\}_{i\in\{1,\ldots,n\}\setminus I_m} \cup \{\tilde{y}_i^s, \tilde{x}_i^s\}_{i\in I_m}$ diverge, $\left\|\hat{\beta}_n^s\right\| \to \infty$ as $s \to \infty$. The rest of the proof is done for weighting schemes with strictly positive weights because the result for $\hat{w}_n^R$ follows from Gervini and Yohai (2000, Theorem 3.3) and Rousseeuw and Leroy (1987).

Since the data are almost surely in a general position and there is at least $p+1$ unmodified points (otherwise $\varepsilon_n^* = 0$ and the theorem holds trivially), we can find for any $\hat{\beta}_n^s$ an observation (y_j, x_j) such that $x_j^\top \hat{\beta}_n^s \neq 0$. Since there is only a finite number of observations, we can assume without loss of generality that this index j of the observation (y_j, x_j) is common to all $\hat{\beta}_n^s, s \in \mathbb{N}$. Since $\left\|\hat{\beta}_n^s\right\| \to \infty$, $\left|x_j^\top \hat{\beta}_n^s\right|$ cannot be bounded for $s \in \mathbb{N}$ unless there is a decomposition $\hat{\beta}_n^s = \hat{\beta}_n^{s,h} + \hat{\beta}_n^{s,r}$ such that $\lim_{s\to\infty} x_j^\top \hat{\beta}_n^{s,h} = 0$ and $\left\|\hat{\beta}_n^{s,r}\right\|$ is bounded for $s \in \mathbb{N}$. In such a case, $x^\top \lim_{n\to\infty} \hat{\beta}_n^{s,h} = 0$ defines a hyperplane, and by the same argument as above, we can find another point $(y_j^{'}, x_j^{'})$ such that $x_j^{'}$ does not belong to this hyperplane anymore, $\left(x_j^{'}\right)^\top \lim_{n\to\infty} \hat{\beta}_n^{s,h} \neq 0$. Hence, $\left\|\hat{\beta}_n^s\right\| \to \infty$ implies that $\left|x_j^\top \hat{\beta}_n^s\right|$ and $\left|r_j^2\left(\hat{\beta}_n^s\right)\right|$ diverge as $s \to \infty$ for some $j \in \{1, \ldots, n\}\setminus I_m$.

We will use this consequence to prove that some $\hat{\beta}_n^s$ cannot correspond to 2S-LWS estimates, which results in contradiction. First, the initial robust estimator has a breakdown point equal to or higher than m/n and the corresponding initial estimates $(\hat{\beta}_n^{0,s}, \hat{\sigma}_n^{0,s}), s \in \mathbb{N}$, based on samples C_m^s are thus bounded: there is some $K > 0$ such that $\max_{s\in\mathbb{N}} \left\{\left\|\hat{\beta}_n^{0,s}\right\|, \left|\hat{\sigma}_n^{0,s}\right|, 1/\left|\hat{\sigma}_n^{0,s}\right|\right\} < K$. Consequently, the residuals $y_i - \hat{\beta}_n^{0,s} x_i$ are bounded as well, $r_i^2(\hat{\beta}_n^{0,s}) < K_r$, for all $s \in \mathbb{N}$ and $i \in \{1, \ldots, n\}\setminus I_m$.

Moreover, as the sample size n is fixed, the weights assigned to the jth observation by weighting functions $\hat{w}_n^Q, \hat{w}_n^D$, and $\hat{w}_n^P$ defined in (10), (11), and (12), respectively, are bounded both from above and below. This is a consequence of the following observations: (i) $(G_n^0)^{-1}(t)$ corresponds to $r_{[tn]}^2\left(\hat{\beta}_n^{0,s}\right)$, which is bounded by K_r at least for $t \leq 1-\varepsilon_n^*$; (ii) $1/K \leq \left|\hat{\sigma}_n^0\right| \leq K$; (iii) $1/(2n) \leq t \leq 1 - 1/(2n)$; and finally, $\hat{g}_n^0\left\{(G_n^0)^{-1}(t)\right\} > 1/(1.06n^{4/5})\mathcal{K}(0)$, where $\mathcal{K}$ denotes a kernel function (which is uniform in our case). Therefore, we can find $\delta_w > 0$ such that the weight assigned to the observation (y_j, x_j) is bounded by δ_w from below and by $1/\delta_w$ from above:

$$\delta_w < \min\left\{\hat{w}_n^Q(t_j), \hat{w}_n^D(t_j), \hat{w}_n^P(t_j), 1/\hat{w}_n^Q(t_j), 1/\hat{w}_n^D(t_j), 1/\hat{w}_n^P(t_j)\right\},$$

where $t_j = G_n\left\{r_j^2\left(\hat{\beta}_n^{0,s}\right)\right\} - \frac{1}{2n}$. An important consequence of this result is that the weighted residual $w\left\{r_j^2\left(\hat{\beta}_n^s\right)\right\} r_j^2\left(\hat{\beta}_n^s\right) > \delta_w r_j^2\left(\hat{\beta}_n^s\right) \to \infty$ diverges for $s \to \infty$ and the same then applies

to the objective function (3) of any considered 2S-LWS estimator.

On the other hand, we can evaluate the 2S-LWS objective function at $\hat{\beta}_n^{0,s}$. Consider a squared residual $r^2 = (y_i - x_i^\top \hat{\beta}_n^{0,s})^2$ corresponding to an arbitrary observation from any modified sample C_m^s and the respective initial estimate $\hat{\beta}_n^{0,s}$; $i = 1, \ldots, n$ and $s \in \mathbb{N}$. Further, let us take $t = G_n(r^2) - \frac{1}{2n} \in \langle \frac{1}{2n}, 1 - \frac{1}{2n} \rangle$, which is valid irrespective of the actual sample C_m^s. Then it holds for the weighted regression residual $w(t)r^2$ using any of the three proposed weight functions (10), (11), and (12) that

$$\begin{aligned}
\hat{w}_n^Q(t)r^2 &\leq K\frac{F_\chi^{-1}\left(1-\frac{1}{2n}\right)}{r^2} I(r^2>0)r^2 \leq K F_\chi^{-1}\left(1-\frac{1}{2n}\right), \\
\hat{w}_n^D(t)r^2 &\leq \frac{f_\chi\left(Kr^2\right)}{(1.06n^{4/5})\mathcal{K}(0)} r^2 \leq f_\chi\left(Kr^2\right) r^2 = \frac{e^{-Kr^2/2}(Kr^2)^{-1/2}}{\sqrt{2}\Gamma(\frac{1}{2})} r^2, \\
\hat{w}_n^P(t)r^2 &\leq \frac{1-F_\chi\left(Kr^2\right)}{1/(2n)} r^2 \leq \left\{1-F_\chi\left(Kr^2\right)\right\} \frac{4n}{K}\frac{Kr^2}{2} \leq \frac{4n}{K}\int_{\frac{1}{2}Kr^2}^{+\infty} t^{-1/2}e^{-t}dt,
\end{aligned}$$

by the definition of f_χ and F_χ. Hence, the weighted residuals at $\hat{\beta}_n^{0,s}$ in the 2S-LWS objective function can be bounded by some $K_w > 0$ uniformly in $r \in \mathbb{R}$ and $s \in \mathbb{N}$ (n is fixed).

Therefore, the 2S-LWS objective function at $\hat{\beta}_n^{0,s}$ is bounded by nK_w for all $s \in \mathbb{N}$, but at the same time, this objective function at $\hat{\beta}_n^s$ exceeds nK_w for some sufficiently large s. This contradicts the assumption that $\hat{\beta}_n^s$ is an 2S-LWS estimate for C_m^s. Consequently, the breakdown point of 2S-LWS must be at least ε_n^*, and asymptotically, $\lim_{n\to\infty} \varepsilon_n^* = \min\{\varepsilon^{0*}, 1/2\} = \varepsilon^{0*}$, which concludes the proof. □

B Proofs of asymptotic properties

Proof of Theorem 5.1: In linear regression models, this result is derived for a general weight function in Mašíček (2004) under more restrictive conditions such as explanatory variables being independent and identically distributed on a bounded support. Here we prove this theorem under more general conditions A1 to A3, but we limit ourselves only to step functions $w = w_s$ to avoid lengthy technical derivations by using the (non)linear-regression results of Čížek (2006). Thus, this proof assumes that the weight function w is a left-continuous step function with steps at points $D = \{d_1, \ldots, d_J\}$ in interval $(0, 1)$ (Assumption A3) and possibly at $d_{J+1} = 1$.

First of all, the objective function of LWS,

$$\begin{aligned} S_n^{(LWS)}(w;\beta) &= \sum_{i=1}^{n} r_i^2(\beta) w \left[G_n\left\{r_i^2(\beta)\right\} - \frac{1}{2n} \right] \\ &= \sum_{j=1}^{J} \{w(d_j) - w(d_{j+1})\} \left[\sum_{i=1}^{n} r_i^2(\beta) I\Big(r_i^2(\beta) \le r_{[d_j n]}^2(\beta)\Big) \right] \qquad (23) \\ &+ w(d_{J+1}) \sum_{i=1}^{n} r_i^2(\beta), \qquad (24) \end{aligned}$$

is actually a weighted sum of the objective functions of some LTS estimators; the sums in the square brackets of (23) are equivalent to the LTS objective functions (2) with trimming constants $h = [d_j n]$ (Čížek, 2006). Therefore, we can now employ the existing asymptotic results for LTS by applying them to every element of sum (23). In this context, note that Assumption A covers all the assumptions relevant for the linear regression model used by Čížek (2006) except for the identification assumption that is verified below.

Next, the LWS estimator, minimizing its objective function $S_n^{(LWS)}(w;\beta)$, can be also obtained from the normal equations $\partial S_n^{(LWS)}(w;\beta)/\partial\beta = 0$. Using expansion (23), Čížek (2006, Lemma 1) implies that the normal equations can almost surely be expressed as

$$\frac{\partial S_n^{(LWS)}(w;\beta)}{\partial\beta} = \sum_{i=1}^{n} r_i(\beta) x_i w \left[G_n\left\{r_i^2(\beta)\right\} - \frac{1}{2n} \right] = 0. \qquad (25)$$

The second derivative of the objective function $\partial^2 S_n^{(LWS)}(w;\beta)/\partial\beta\partial\beta^\top$ can be analogously expressed as

$$\frac{\partial^2 S_n^{(LWS)}(w;\beta)}{\partial\beta\partial\beta^\top} = \sum_{i=1}^{n} x_i x_i^\top w \left[G_n\left\{r_i^2(\beta)\right\} - \frac{1}{2n} \right].$$

Moreover, Assumption A allows us to use the result of Čížek (2006, Lemma 3), which implies uniformly in β (over any compact subset of $\mathbb{R}^p$) that

$$\frac{1}{n}\frac{\partial S_n^{(LWS)}(w;\beta)}{\partial\beta} \to \mathrm{E}\left\{ r_i(\beta) x_i w \left[G_\beta\left\{r_i^2(\beta)\right\}\right]\right\} = S'(\beta) \qquad (26)$$

and

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top w \left[G_n\left\{r_i^2(\beta)\right\} - \frac{1}{2n} \right] \to \mathrm{E}\left\{ x_i x_i^\top w \left[G_\beta\left\{r_i^2(\beta)\right\}\right]\right\} = Q(\beta)$$

in probability for $n \to \infty$, where G_β denotes the distribution function of $r_i^2(\beta)$. The matrix $Q(\beta)$ is a positive semidefinite matrix, and at β^0, the matrix $Q(\beta^0) = Q$ is a nonsingular positive definite matrix (Assumption A1), which guarantees that the identification assumption of Čížek (2006) is satisfied. (Note that a general proof of the identification conditions under trimming without using derivatives of the objective function is given in Čížek, 2007.) Due to the continuity of $Q(\beta)$ at β^0 and the uniform convergence of $Q_n(\beta)$ to $Q(\beta)$, it is possible to find some $n_0 \in \mathbb{N}$ such that the matrix $Q_n(\beta)$ is positive definite in a neighborhood of β^0 (and positive semidefinite elsewhere) with a probability greater than $1-\varepsilon$ for any $\varepsilon > 0$ and $n > n_0$.

For a sufficiently large n, we will now show that there is a solution to the normal equations (25) in an arbitrarily small neighborhood of β^0. Because $S'(\beta^0) = 0$, see (26) and Assumption A2, and $Q_n(\beta)$ is positive definite around β^0 and positive semidefinite elsewhere, this solution is unique (with an arbitrarily high probability) and equals the LWS estimate minimizing $S_n^{(LWS)}(w;\beta)$. To find the solution of (25), the asymptotic linearity of LTS is employed in a neighborhood $U(\beta^0, n^{-\frac{1}{2}}M)$ of β^0, where $M > 0$. To characterize $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$, one can express it as $\beta = \beta^0 - n^{-\frac{1}{2}}t$ for $t \in T_M = \{t : \|t\| \leq M\}$. Thus, using the asymptotic linearity theorem for LTS (Čížek, 2006; Theorem 1) and the expansion (23)–(24) of the LWS objective function, we can write that

$$\frac{\partial S_n^{(LWS)}(w;\beta^0 - n^{-\frac{1}{2}}t)}{\partial\beta} = \frac{\partial S_n^{(LWS)}(w;\beta^0)}{\partial\beta} - n^{\frac{1}{2}}Qt \cdot C(w) + o_p(1) \tag{27}$$

uniformly for all $t \in T_M$ and $M > 0$, where

$$C(w) = \sum_{j=1}^{J} \{w(d_j) - w(d_{j+1})\} \{d_j - 2q_j f(q_j)\} + w(d_{J+1}) \tag{28}$$

and notation $q_j = \sqrt{G^{-1}(d_j)}$ is used for $j = 1, \ldots, J$.

Thus, we have to show that, with an arbitrarily high probability, there is a $t_n^* \in T_M$ such that $\beta^0 - n^{-\frac{1}{2}}t_n^*$ solves the normal equations $\partial S_n^{(LWS)}(w;\beta^0 - n^{-\frac{1}{2}}t_n^*)/\partial\beta = 0$. At such a solution t_n^*, equation (27) implies

$$\frac{\partial S_n^{(LWS)}(w;\beta^0)}{\partial\beta} = n^{\frac{1}{2}}QC(w) \cdot t_n^* + o_p(1) \tag{29}$$

and, recalling that Q is a nonsingular matrix and $C(w) \neq 0$ (see (32)–(33)),

$$t_n^* = Q^{-1}C(w)^{-1} \cdot \frac{1}{\sqrt{n}} \frac{\partial S_n^{(LWS)}(w;\beta^0)}{\partial \beta} + o_p\left(n^{-\frac{1}{2}}\right) \tag{30}$$

as $n \to \infty$. To prove that t_n^* is bounded in probability, we have to show that

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial S_n^{(LWS)}(w;\beta^0)}{\partial \beta} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} r_i(\beta^0) x_i w \left[G_n\left\{ r_i^2\left(\beta^0\right)\right\} - \frac{1}{2n} \right] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} r_i(\beta^0) x_i w \left[G\left\{ r_i^2\left(\beta^0\right)\right\} \right] + o_p(1) \end{aligned} \tag{31}$$

is bounded in probability (equality (31) follows, after using expansion (23)–(24), from Čížek, 2006, Theorem 4 and its proof). Due to decomposition (23)–(24), equation (31) can be rewritten as a finite sum of random variables that are all asymptotically normally distributed (Čížek, 2006; Theorem 4). Hence, (31) and t_n^* in (30) are bounded in probability, and for some $n_0 \in \mathbb{N}$ and $\varepsilon > 0$, the left-hand side of (27) equals zero for some $t_n^* \in T_M, n > n_0$, with probability higher than $1-\varepsilon$. Then $\beta^0 - n^{-\frac{1}{2}} t_n^*$ is the unique solution of (25), and consequently, the LWS estimate itself, $\hat{\beta}_n^{(LWS)} = \beta^0 - n^{-\frac{1}{2}} t_n^*$. Apparently, it holds that $\sqrt{n}\left(\hat{\beta}_n^{(LWS)} - \beta^0\right) = t_n^* = \mathcal{O}_p(1)$, which implies the $\sqrt{n}$-consistency of LWS.

Finally, we have to prove the asymptotic normality of LWS, that is, to find the asymptotic distribution of t_n^*. Because $C(w)$ and Q in (30) are constants, we just have to derive the asymptotic distribution of (31). The summands of (31),

$$r_i(\beta^0) x_i^\top w \left[G\left\{ r_i^2\left(\beta^0\right)\right\} \right] \equiv \varepsilon_i x_i^\top w \left\{ G(\varepsilon_i^2) \right\},$$

form by Assumption A2 a sequence of martingale differences with finite variances since the law of large numbers for L^1-mixingales (Andrews, 1988) implies

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 x_i x_i^\top w^2 \left\{ G(\varepsilon_i^2) \right\} \to \operatorname{var}\left[\varepsilon_1 x_1 w \left\{ G(\varepsilon_1^2) \right\} \right]$$

in probability as $n \to \infty$. Hence, we can employ the central limit theorem for martingale differences (e.g., Davidson, 1994, Theorem 24.3) for (31), which proves its asymptotic normality.

By (30) and $\sqrt{n}\left(\hat{\beta}_n^{(LWS)} - \beta^0\right) = t_n^*$, it follows that

$$\sqrt{n}\left(\hat{\beta}_n^{(LWS)} - \beta^0\right) \xrightarrow{\mathcal{L}} N(0, V_w),$$

where $V_w = C(w)^{-2}Q^{-1}\text{var}\left[\varepsilon_1 x_1 w\left\{G(\varepsilon_i^2)\right\}\right]Q^{-1}$.

The result of Theorem 5.1 then follows by rewriting of the constant $C(w)$ in (28): using integration by parts,

$$\begin{aligned} 2q_j f(q_j) - d_j &= \{q_j f(q_j) + q_j f(-q_j)\} - \{F(q_j) - F(-q_j)\} && (32) \\ &= [\varepsilon f(\varepsilon)]_{-q_j}^{q_j} - \int_{-q_j}^{q_j} f(\varepsilon)d\varepsilon = \int_{-q_j}^{q_j} \varepsilon f'(\varepsilon)d\varepsilon, && (33) \end{aligned}$$

and $\int_{-\infty}^{+\infty} \varepsilon f'(\varepsilon)d\varepsilon = -1$ for $d_j \to 1$ and $q_j \to \infty$, (28) can be expressed as

$$\begin{aligned} C(w) &= -\sum_{j=1}^{J}\{w(d_j) - w(d_{j+1})\}\int_{-q_j}^{q_j} \varepsilon f'(\varepsilon)d\varepsilon + w(d_{J+1}) \\ &= -\sum_{j=1}^{J}\int_{-q_j}^{q_j}\{w(d_j) - w(d_{j+1})\}\varepsilon f'(\varepsilon)d\varepsilon - \int_{-\infty}^{+\infty} w(d_{j+1})\varepsilon f'(\varepsilon)d\varepsilon \\ &= -\int_{-\infty}^{+\infty} \varepsilon w\left\{G(\varepsilon^2)\right\} f'(\varepsilon)d\varepsilon, \end{aligned}$$

which concludes the proof. □

Proof of Corollary 5.2: Similarly to the proof of Theorem 5.1, we prove the result for stepwise weight functions $w = w_s$ under Assumption A. In linear regression models with independent and identically distributed variables, one can derive the result for a general weight function analogously using the asymptotic linearity of the LWS normal equation derived in Víšek (2002a) or Mašíček (2004), who also use the representation (23)–(24).

First note that the first part of the proof of Theorem 5.1 holds also for estimated stepwise weight function since it only employs the asymptotic linearity for LTS and multiplies the corresponding expressions by appropriate weights. Hence, expressions (27)–(31) hold also for a weight function $\hat{w}_n$. To prove the claim of the corollary, we just have to show that the difference between (29) with the weight function w and

$$\frac{1}{\sqrt{n}}\frac{\partial S_n^{(LWS)}(\hat{w}_n; \beta^0)}{\partial \beta} = QC(\hat{w}_n)\cdot t_n^* + o_p(1)$$

is negligible in probability. Because $|C(\hat{w}_n) - C(w)| \to 0$ as $n \to \infty$ follows from the consistency of $\hat{w}_n, \hat{w}_n \to w$, and continuity of G, G^{-1}, and f (use Assumption A2 and $G(z^2) = F(|z|) - F(-|z|)$), we only have to show that

$$\frac{1}{\sqrt{n}} \frac{\partial S_n^{(LWS)}(\hat{w}_n; \beta^0)}{\partial \beta} - \frac{1}{\sqrt{n}} \frac{\partial S_n^{(LWS)}(w; \beta^0)}{\partial \beta} = o_p(1). \tag{34}$$

Using (31), we can rewrite this difference up to a term negligible in probability as

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} r_i(\beta^0) x_i \left\{ \hat{w}_n \left[G\left\{ r_i^2(\beta^0) \right\} \right] - w \left[G\left\{ r_i^2(\beta^0) \right\} \right] \right\} \\ \leq \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i \left\{ \hat{w}_n \left[G(\varepsilon_i^2) \right] - w \left[G(\varepsilon_i^2) \right] \right\} I\big(1/K \leq \varepsilon_i^2 \leq K\big) \qquad (35) \\ + \; & \frac{C}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i I\big(\varepsilon_i^2 \notin \langle 1/K, K \rangle\big) \qquad (36) \end{aligned}$$

for any $K > 1$ and some $C > 0$ (weight functions are bounded).

First, (36) has a zero expectation (Assumption A2) and variance equal to $V_2(K) = \mathrm{E}\{\varepsilon_i^2 I(\varepsilon_i^2 \notin \langle 1/K, K \rangle) x_i x_i^\top\}$. Since $V_2(K) \to 0$ as $K \to \infty$, the Chebyshev inequality implies that (36) can be made arbitrarily small in probability by choosing sufficiently large K.

Next, for a given K, let us denote $\nu_i(K)$ the summands in (35),

$$\nu_i(K) = \varepsilon_i x_i \left\{ \hat{w}_n \left[G(\varepsilon_i^2) \right] - w \left[G(\varepsilon_i^2) \right] \right\} I\big(1/K \leq \varepsilon_i^2 \leq K\big).$$

The conditional expectations of these summands are zero,

$$\mathrm{E}\left[\nu_i(K) \middle| \, \varepsilon_1, \ldots, \varepsilon_{i-1}, x_1, \ldots, x_{i-1} \right] = 0,$$

and due to the n^α consistency of $\hat{w}_n$, the variance of $n^\alpha \nu_i(K)$ is bounded since by the Schwarz inequality

$$\begin{aligned} & \mathrm{E}\left[\varepsilon_i^2 n^{2\alpha} \left\{ \hat{w}_n \left[G(\varepsilon_i^2) \right] - w \left[G(\varepsilon_i^2) \right] \right\}^2 I\big(1/K \leq \varepsilon_i^2 \leq K\big) x_i x_i^\top \right] \\ & \leq \mathrm{E}\left[\varepsilon_i^2 x_i x_i^\top \right] \cdot \mathrm{E}\left[n^{2\alpha} \left\{ \hat{w}_n \left[G(\varepsilon_i^2) \right] - w \left[G(\varepsilon_i^2) \right] \right\}^2 I\big(1/K \leq \varepsilon_i^2 \leq K\big) \right] \end{aligned}$$

(note that the convergence of $\hat{w}_n \to w$ in probability implies convergence in mean because all

weight functions are bounded). Hence, we can apply the law of large numbers for L^2-mixingales (Davidson and de Jong, 1997, Corollary 2.1) to (35) written as $n^{-1/2-1/\alpha}\sum_{i=1}^{n} n^{\alpha}\nu_i(K) \to 0$ in probability as $n \to \infty$.

Consequently by letting $K \to \infty$, both (35) and (36) are negligible in probability and (34) holds. □

Proof of Lemma 5.3: In the case of weights $\hat{w}_n^R$, the result follows from Gervini and Yohai (2002, Lemma 4.1). For other weight functions, $\hat{w}_n^Q, \hat{w}_n^D, \hat{w}_n^P$, the continuity and differentiability of F_χ and f_χ, which implies their uniform continuity on any compact subset of their support, indicates that it is enough to prove for any $0 < a < b < 1$ that

$$\sup_{z\in\mathbb{R}} |G_n(z) - G(z)| = \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big), \tag{37}$$

$$\sup_{t\in\langle a,b\rangle} |G_n^{-1}(t) - G^{-1}(t)| = \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big), \tag{38}$$

$$\sup_{z\in\mathbb{R}} |\hat{g}_n^0(z) - g(z)| = \mathcal{O}_p\Big(n^{-\frac{1}{2}}h_n^{-1}\Big), \tag{39}$$

where h_n denotes the bandwidth used for estimating $\hat{g}_n^0$. The statement (37) follows from Gervini and Yohai (2002, Lemma 4.2) because $G(z) = F^+(\sqrt{z})$, where F^+ denotes the distribution function of $|\varepsilon_i|$. Further, (39) is derived in Cai and Roussas (1992), for instance.

Hence, we only have to prove that (37) implies (38). For a given $t \in \langle a, b\rangle$, let $z = G_n^{-1}(t)$. The Taylor expansion implies

$$G^{-1}(t) = G^{-1}\{G(z)\} + \frac{t - G(z)}{g(\xi)} \le G_n^{-1}(t) + \frac{1}{g(\xi)}\left\{|G_n(z) - G(z)| + \frac{1}{n}\right\},$$
$$G^{-1}(t) = G^{-1}\{G(z)\} + \frac{t - G(z)}{g(\xi)} \ge G_n^{-1}(t) + \frac{1}{g(\xi)}\left\{|G_n(z) - G(z)| - \frac{1}{n}\right\},$$

where $\xi \in (G(z), t)$. Because $|G_n(z) - G(z)| = \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)$ and $g(\xi) > c_g > 0$ for any $\xi \in \langle a/2, (1+b)/2\rangle$ by Assumption A2,

$$|G_n^{-1}(t) - G^{-1}(t)| \le c_g^{-1}\left\{|G_n(z) - G(z)| + \frac{1}{n}\right\}$$

with an arbitrarily high probability and (37) thus implies (39). □

Proof of Theorem 5.4: The consistency of the proposed variance-matrix estimator is a direct consequence of the weak law of large number; we use here its form for L^1-mixingales due to

Andrews (1988).

First, Assumption A1 indicates that $\{x_i\}_{i=1}^n$ is a β-mixing sequence with finite rth moments, $r > 2$. Hence, $\hat{Q}_n = \sum_{i=1}^n x_i x_i^\top / n \to Q$ in probability as $n \to \infty$. Similarly, since $w(t)$ is bounded, $w(t) \leq K_w$ for $t \in \langle 0, 1\rangle$, and Theorem 5.1 and Corollary 5.2 imply $|e_{in} - \varepsilon_i| = \mathcal{O}_p\left(n^{-\frac{1}{2}}\right)$ (where $e_{in} = r_i(\hat{\beta}_n)$), Assumption A2 ensures that $x_i e_{in} w[G_n(e_{in}^2)]$ possesses finite rth moments ($\left\|x_i e_{in} w[G_n(e_{in}^2)]\right\| \leq \|x_i \varepsilon_i\| K_w + \mathcal{O}_p\left(n^{-\frac{1}{2}}\right)$). Due to the symmetry of the LWS objective function with respect to regression residuals and the symmetry of the error distribution F, $\left\{x_i e_{in} w[G_n(e_{in}^2)]\right\}_{i=1}^n$ forms a sequence of martingale differences (see Assumption A2 and also the proof of Theorem 5.1). Hence, by the law of large numbers for triangular arrays (Andrews, 1988)

$$\hat{C}_{Vn} = \frac{1}{n}\sum_{i=1}^n x_i^2 e_{in}^2 w^2\left\{G_n(e_{in}^2) - \frac{1}{2n}\right\} \to \text{var}\left[x_1 \varepsilon_1 w\{G(\varepsilon_1^2)\}\right],$$

since $x_1 e_{1n} w\{G(e_{1n}^2)\} \to x_1 \varepsilon_1 w\{G(\varepsilon_1^2)\}$ in probability as $n \to \infty$ (see Assumption A3, decomposition (23)–(24), and Čížek, 2006, Lemma 3).

Next, we estimate the denominator of the LWS covariance matrix,

$$-\int \varepsilon w\{G(\varepsilon^2)\} f'(\varepsilon) d\varepsilon = -\int \varepsilon\left[w_s\{G(\varepsilon^2)\} + w_c\{G(\varepsilon^2)\}\right] f'(\varepsilon) d\varepsilon = C_I^S + C_I^C.$$

Let us first prove that the latter term C_I^C can be consistently estimated by (21). By integration by parts,

$$\begin{aligned} -\int \varepsilon w_c\{G(\varepsilon^2)\} f'(\varepsilon) d\varepsilon &= \int w_c\{G(\varepsilon^2)\} f(\varepsilon) d\varepsilon + \int 2\varepsilon^2 w_c'\{G(\varepsilon^2)\} g(\varepsilon^2) f(\varepsilon) d\varepsilon \\ &= \text{E}[w_c\{G(\varepsilon_1^2)\}] + 2\text{E}[\varepsilon_1^2 w_c'\{G(\varepsilon_1^2)\} g(\varepsilon_1^2)], \end{aligned}$$

because the error term ε_i has finite second moments (Assumption A2) and w_c is bounded and continuously differentiable (Assumption A3). Since the proof follows the same steps for both part, we prove only the consistency of the first term. We assume that $\hat{G}_n$ and $\hat{g}_n$ are uniformly consistent estimators of the distribution and density functions, G and g, respectively. That is, for any $\delta > 0$ there is n_0 such that $\sup_z |\hat{G}_n(z) - G(z)| + \frac{1}{2n} < \delta$ and $\sup_z |\hat{g}_n(z) - g(z)| < \delta$ for $n > n_0$ with probability larger than $1 - \varepsilon$. Due to the continuity of w_c on $\langle 0, 1\rangle$, w_c is uniformly continuous and $\sup_{|t-t'|<\delta} |w_c(t) - w_c(t')| < \eta$ for all $t, t' \in \langle 0, 1\rangle$ and some $\eta > 0$, where η can

be made arbitrarily small by choosing a sufficiently small δ. Consequently, we can write for $n \to \infty$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w_c\left\{\hat{G}_n(e_{in}^2) - \frac{1}{2n}\right\} &= \frac{1}{n}\sum_{i=1}^{n} w_c\left\{G(e_{in}^2)\right\} + \frac{1}{n}\sum_{i=1}^{n}\left[w_c\left\{G(e_{in}^2)\right\} - w_c\left\{\hat{G}_n(e_{in}^2) - \frac{1}{2n}\right\}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} w_c\left\{G(e_{in}^2)\right\} + o_p(1)\end{aligned}$$

As $e_{in} \to \varepsilon_i$ in probability as $n \to \infty$ and functions w_c and G are continuous, the Slutsky lemma and the law of large numbers for triangular arrays imply $\frac{1}{n}\sum_{i=1}^{n} w_c\left\{G(e_i^2)\right\} \to \mathrm{E}\left[w_c\left\{G(\varepsilon_1^2)\right\}\right]$, and thus, $\hat{C}_{In}^C \to C_I^C$ in probability as $n \to \infty$.

Finally, let us deal with C_I^S, which is claimed to be consistently estimated by (20). Let q_j^2 denote $G^{-1}(d_j)$ for $j = 1, \ldots, J$. Using decomposition (23)–(24) for $w_s, w_s(1) = 0$, integration by parts leads to

$$\begin{aligned}\int \varepsilon w_s\{G(\varepsilon^2)\} f'(\varepsilon) d\varepsilon &= \sum_{j=1}^{J}\{w_s(d_j) - w_s(d_{j+1})\}\int_{-q_j}^{q_j} \varepsilon f'(\varepsilon) d\varepsilon = \\ &= \sum_{j=1}^{J}\{w(d_j) - w(d_{j+1})\}\left\{[\varepsilon f(\varepsilon)]_{-q_j}^{q_j} - \int_{-q_j}^{q_j} f(\varepsilon) d\varepsilon\right\} \\ &= \sum_{j=1}^{J}\{w(d_j) - w(d_{j+1})\} \times \\ &\qquad \times \left[q_j\{f(q_j) + f(-q_j)\} - \{F(q_j) - F(-q_j)\}\right] \\ &= \sum_{j=1}^{J}\{w(d_j) - w(d_{j+1})\}\left[2q_j^2 g(q_j^2) - d_j\right]\end{aligned}$$

because $G(z^2) = F(z) - F(-z)$ and $g(z^2) = \{f(z) + f(-z)\}/2z$. The only unknown quantities are quantiles q_j^2, which are estimated by $\hat{q}_{jn}^2 = e_{[d_j n]}^2$, and density g, which is estimated by $\hat{g}_n$. Since $\hat{q}_{jn}^2 \to q_j^2$ in probability as $n \to \infty$ (Čížek, 2004, Lemma A.2) and $\hat{g}_n(q_j)$ is uniformly consistent and bounded, it holds that

$$\hat{q}_j^2 \hat{g}_n(\hat{q}_j^2) - q_j^2 g(q_j^2) = (\hat{q}_j^2 - q_j^2)\hat{g}_n(\hat{q}_j^2) - q_j^2\left\{\hat{g}_n(\hat{q}_j^2) - g(q_j^2)\right\} \to 0$$

in probability as $n \to \infty$, which closes the proof. □

References

[1] Andrews, D. W. K. (1988) Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* **4**, 458–467.

[2] Andrews, D. W. K. (1993) An introduction to econometric applications of empirical process theory for dependent random variables. *Econometric Reviews* **12**, 183–216.

[3] Balke, N. S. & T. B. Fomby (1994) Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. *Journal of Applied Econometrics* **9**, 181–200.

[4] Cai, Z. & G. G. Roussas (1992) Uniform strong estimation under α-mixing with rates. *Statistics & Probability Letters* **15**, 47–55.

[5] Castellana, J. V. & M. R. Leadbetter (1986) On smoothed probability density estimation for stationary processes. *Stochastic Processes and their Applications* **21**, 179–193.

[6] Čížek, P. (2004) Asymptotics of least trimmed squares regression. CentER discussion paper 2004/72, CentER, Tilburg University.

[7] Čížek, P. (2006) Least trimmed squares under dependence. *Journal of Statistical Planning and Inference* **136**, 3967–3988.

[8] Čížek, P. (2007) General trimmed estimation: robust approach to nonlinear and limited dependent variable models. CentER discussion paper 2007/1, CentER, Tilburg University, submitted to Econometric Theory.

[9] Croux, C., Rousseeuw, P. and Hössjer, O. (1994) Generalized S-estimators. *Journal of the American Statistical Association* **89**, 1271–1281.

[10] Davidson, J. (1994) *Stochastic Limit Theory*. New York: Oxford University Press.

[11] Davies, L. (1990) The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics* **18**, 1651–1675.

[12] Einmahl, U. & D. M. Mason (2005) Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* **33**, 1380–1403.

[13] Engler, E. & B. Nielsen (2007) The empirical process of autoregressive residuals. Discussion Paper, Nutffield College, University of Oxford, `http://www.nuff.ox.ac.uk/Economics/papers/2007/w1/ EnglerNielsen07.pdf`.

[14] Genton, M. G. & A. Lucas (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society, Series B* **65**, 81–94.

[15] Gervini, D. & V. J. Yohai (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics* **30**, 583–616.

[16] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw & W. A. Stahel (1986) *Robust statistics: The approach based on influence function*. New York: Wiley.

[17] He, X. & S. Portnoy (1992) Reweighted LS estimators converge at the same rate as the initial estimator. *The Annals of Statistics* **20**, 2161–2167.

[18] Hössjer, O. (1992) On the optimality of S-estimators. *Statistics and Probability Letters* **14**, 413–419.

[19] Hubert, M. & P. J. Rousseeuw (1997) Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference* **57**, 153–163.

[20] Jurečková, J. (1984) Regression quantiles and trimmed least squares estimator under a general design. *Kybernetika* **20**, 345–357.

[21] Mašíček, L. (2004) Diagnostics and sensitivity of robust models. Unpublished Ph.D. Thesis, Faculty of Mathematics and Physics, Charles University, Prague.

[22] Mokkadem, A. (1988) Mixing properties of ARMA processes. *Stochastic Processes and Their Application* **29**, 309–315.

[23] Pagan, A. and Ullah, A. (1999) *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

[24] Preminger, A. & R. Franck (2007) Foreign exchange rates: a robust regression approach. *International Journal of Forecasting* **23**, 71–84.

[25] Rousseeuw, P. J. (1984) Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880.

[26] Rousseeuw, P. J. (1985) Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze & W. Wertz (eds.) *Mathematical statistics and applications, Vol. B*. Dordrecht: Reidel, pp. 283–297.

[27] Rousseeuw, P. J. (1997) Introduction to positive-breakdown methods. In G. S. Maddala, & C. R. Rao (eds.) *Handbook of statistics 15: Robust inference*. Amsterdam: Elsevier, pp. 101–121.

[28] Rousseeuw, P. J. & C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**, 1273–1283.

[29] Rousseeuw, P. J. & A. M. Leroy (1987). *Robust regression and outlier detection*. New York: Wiley.

[30] Rousseeuw, P. J. & V. J. Yohai (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle & R. D. Martin (eds.) *Robust and nonlinear time series analysis, Lecture notes in statistics, Vol. 26*. Springer, New York, 256–272.

[31] Sakata, S. & H. White (1998) High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. *Econometrica* **66**, 529–567.

[32] Sakata, S. & H. White (2001) S-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics* **103**, 5–72.

[33] Simpson, D. G., D. Ruppert & R. J. Carroll (1992) On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* **87**, 439–450.

[34] Stromberg, A. J., O. Hössjer & D. M. Hawkins (2000) The least trimmed difference regression estimator and alternatives. *Journal of the American Statistical Association* **95**, 853–864.

[35] Tableman, M. (1994) The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators. *Statistics & Probability Letters* **19**, 329–337.

[36] Temple, J. R. W. (1998) Robustness tests of the augmented Solow model. *Journal of Applied Econometrics* **13**, 361–375.

[37] Van Dijk, D., P. H. Franses & A. Lucas (1999) Testing for ARCH in the presence of additive outliers. *Journal of Applied Econometrics* **14**, 539–562.

[38] Víšek, J. Á. (2002a) The least weighted squares I. The asymptotic linearity of normal equations. *Bulletin of the Czech Econometric Society* **9**(15), 31–58.

[39] Víšek, J. Á. (2002b) The least weighted squares II. Consistency and asymptotic normality. *Bulletin of the Czech Econometric Society* **9**(16), 1–28.

[40] Víšek, J. Á. (2006) Instrumental weighted variables. *Austrain Journal of Statistics* **35**, 379–387.

[41] Welsh, A. H. & E. Ronchetti (2002) A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* **103**, 287–310.

[42] Woo, J. (2003) Economic, political, and institutional determinants of public deficits. *Journal of Public Economics* **87**, 387–426.

[43] Yohai, V. J. & R. H. Zamar (1988) High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* **83**, 406–413.

[44] Zinde-Walsh, V. (2002) Asymptotic theory for some high breakdown point estimators. *Econometric Theory* **18**, 1172–1196.